Improved Algorithm for Adversarial Linear Mixture MDPs with Bandit Feedback and Unknown Transition

Long-Fei Li, Peng Zhao, Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology, Nanjing University, China
School of Artificial Intelligence, Nanjing University, China
{lilf, zhaop, zhouzh}@lamda.nju.edu.cn

Abstract

We study reinforcement learning with linear function approximation, unknown transition, and adversarial losses in the bandit feedback setting. Specifically, we focus on linear mixture MDPs whose transition kernel is a linear mixture model. We propose a new algorithm that attains an $\widetilde{\mathcal{O}}(d\sqrt{HS^3K} + \sqrt{HSAK})$ regret with high probability, where d is the dimension of feature mappings, S is the size of state space, A is the size of action space, H is the episode length and K is the number of episodes. Our result strictly improves the previous best-known $\widetilde{\mathcal{O}}(dS^2\sqrt{K} + \sqrt{HSAK})$ result in Zhao et al. (2023a) since $H \leq S$ holds by the layered MDP structure. Our advancements are primarily attributed to (i) a new least square estimator for the transition parameter that leverages the visit information of all states, as opposed to only one state in prior work, and (ii) a new self-normalized concentration tailored specifically to handle non-independent noises, originally proposed in the dynamic assortment area and firstly applied in reinforcement learning to handle correlations between different states.

1 INTRODUCTION

Reinforcement Learning (RL) studies the problem where a learner interacts with the environment sequentially and aims to improve the strategy over time. RL has achieved great success in the fields of games (Mnih et al., 2013), robotic control (Schulman et al., 2017), large language models (OpenAI, 2023) and so on.

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

One of the most popular models to describe the RL problem is the Markov Decision Process (MDP) (Puterman, 1994). Significant advances have emerged in learning MDPs with fixed or stochastic loss functions (Jaksch et al., 2010; Azar et al., 2017), however, in many real-world applications, the losses may not be fixed or sampled from certain underlying distributions. As such, the pioneering works of Even-Dar et al. (2009) and Yu et al. (2009) make the first step to formulate and study *adversarial* MDPs, where the loss functions can be chosen adversarially and may change arbitrarily between each time step. Subsequently, many works explore different settings depending on the knowledge of the transition and the type of feedback received, whether it is full-information or bandit feedback (Zimin and Neu, 2013; Rosenberg and Mansour, 2019a,b; Jin et al., 2020a). More detailed discussions are presented in Section 2.

Most existing works studying adversarial MDPs focus on the tabular setting, where the state and action space are small. Yet, in many problems, the state and action space can be large or even infinite. To overcome this challenge, a widely used approach in the literature is *function approximation*, which reparameterizes the action-value function as a function over some feature mapping that maps the state and action to a low-dimensional space. In particular, linear function approximation has gained significant attention (Jin et al., 2020b; Ayoub et al., 2020; Zhou et al., 2021; Li et al., 2023). Amongst these works, linear mixture MDPs (Ayoub et al., 2020) and linear MDPs (Jin et al., 2020b) are two of the most popular models. In this work, we focus on linear mixture MDPs whose transition is a linear mixture model.

The exploration of adversarial linear mixture MDPs remains an emerging area of research. In particular, Cai et al. (2020) first study adversarial linear mixture MDPs with the unknown transition and full-information feedback. They propose a policy optimization algorithm OPPO that achieves $\widetilde{\mathcal{O}}(dH^2\sqrt{K})$

Table 1: Comparisons of regret bounds for adversarial tabular MDPs and linear mixture MDPs with bandit feedback and unknown transition in the literature. S is the size of state space, A is the size of action space, K is the number of episodes and H is the length of each episode, d is the dimension of feature mapping.

	Reference	**Model**	**Regret**
Upper bound	Jin et al. (2020a)	Tabular MDPs	$\widetilde{\mathcal{O}}(HS\sqrt{AK})$
	Zhao et al. (2023a)	Linear Mixture MDPs	$\widetilde{\mathcal{O}}(dS^2\sqrt{K} + \sqrt{HSAK})$
	This work	Linear Mixture MDPs	$\widetilde{\mathcal{O}}(d\sqrt{HS^3K} + \sqrt{HSAK})$
Lower bound	Jin et al. (2018)	Tabular MDPs	$\Omega(H\sqrt{SAK})$
	Zhao et al. (2023a)	Linear Mixture MDPs	$\Omega(dH\sqrt{K} + \sqrt{HSAK})$

regret. The subsequent work by He et al. (2022) enhances the result to $\widetilde{\mathcal{O}}(dH^{3/2}\sqrt{K})$ and shows it is minimax optimal. For the more challenging setting with unknown transition and bandit feedback, Zhao et al. (2023a) achieve an $\widetilde{\mathcal{O}}(dS^2\sqrt{K} + \sqrt{HSAK})$ regret, which exhibits a notable gap to the $\Omega(dH\sqrt{K} + \sqrt{HSAK})$ lower bound established therein.

In this work, we study adversarial linear mixture MDPs with bandit feedback and unknown transition. We strictly improve the result of Zhao et al. (2023a) and make a step towards closing the gap between the upper and lower bound. Specifically, we propose an algorithm that attains $\widetilde{\mathcal{O}}(d\sqrt{HS^3K} + \sqrt{HSAK})$ regret, strictly improving the $\widetilde{\mathcal{O}}(dS^2\sqrt{K} + \sqrt{HSAK})$ regret of Zhao et al. (2023a) since $H \leq S$ by the layered MDP structure. As a byproduct, our result improves the best-known $\widetilde{\mathcal{O}}(HS\sqrt{AK})$ regret of Jin et al. (2020a) for tabular MDPs when $d \leq \sqrt{HA/S}$. We note that though the dependence S of our result is suboptimal, to the best of our knowledge, for this challenging unknown transition and bandit feedback setting, closing the gap regarding the dependence on S for the tabular case is also an open problem. Table 1 summarizes our result and previous related results.

Our algorithm is similar to that of Zhao et al. (2023a): we first estimate the unknown transition parameter and construct corresponding confident sets. Then we apply Online Mirror Descent (OMD) over the occupancy measure space induced by the estimated transition. The most natural approach to estimate the unknown parameter is solving a linear regression problem with the visit statuses of the next states being the target. However, since the learner only visits one state in each step, the visit statuses across different states are no longer independent. This makes the key self-normalized concentration in (Abbasi-Yadkori et al., 2011, Theorem 1), as restated in Lemma 15, not applicable. To address this issue, Zhao et al. (2023a) propose to leverage the transition information of only one state with the largest uncertainty. Though this technique effectively bypasses the issue of state correlation and serves as the first solution for this problem, unfortunately, it discards the visit information of other states, leading to a notable gap to the lower bound.

To enhance the utility of visitation data, we introduce a new least square estimator for the unknown transition parameter that leverages the visit information of *all states*, as opposed to only a single state in Zhao et al. (2023a). As stated before, the noises now are *non-independent* across different states. We address this key challenge by introducing a new self-normalized concentration lemma tailored specifically to accommodate non-independent random noises. This lemma was originally proposed by Périvier and Goyal (2022) for the *dynamic assortment* problem, where a seller selects the subset of products to present to the customer who will then purchase at most one *single* item. They use this lemma to manage the product correlations and we make adaptations to handle the state correlations. This enhancement empowers our algorithm to explore the orientations of every state simultaneously, distinguishing our method from the singular direction approach of Zhao et al. (2023a), and resulting in a tighter bound. To the best of our knowledge, this is the first work that bridges the two distinct fields: dynamic assortment and RL theory. Our innovative use of techniques from dynamic assortment problems to mitigate estimation errors in RL theory is novel and may provide helpful insights for future research.

Organization. The rest of the paper is organized as follows. We first discuss the related work in Section 2 and formulate the problem setup in Section 3. We introduce the proposed algorithm in Section 4 and present the regret guarantee in Section 5. Finally, We conclude the paper in Section 6. Due to page limits, we defer all the proofs to the appendices.

Notation. We denote by $[n]$ the set $\{1, \ldots, n\}$ and use $\mathbb{I}\{\cdot\}$ to denote the indicator function. For a vector $x \in \mathbb{R}^d$ and a positive semi-definite matrix $\Sigma \in \mathbb{R}^{d \times d}$, let $\|x\|_\Sigma = \sqrt{x^\top \Sigma x}$. Let $a \wedge b = \min\{a, b\}$ for all $a, b \in \mathbb{R}$. The $\widetilde{\mathcal{O}}(\cdot)$-notation hides all logarithmic factors.

2 RELATED WORK

In this part, we review related works in the literature.

RL with adversarial losses. Learning tabular RL with adversarial losses has been well-studied in the literature (Even-Dar et al., 2009; Yu et al., 2009; Zimin and Neu, 2013; Rosenberg and Mansour, 2019a,b; Jin et al., 2020a; Shani et al., 2020; Luo et al., 2021). In general, these studies can be divided into two categories based on the type of the algorithm. The first category solves adversarial MDPs using policy-optimization-based methods. The pioneering works of Even-Dar et al. (2009) and Yu et al. (2009) first study adversarial MDPs under the known transition and full-information setting. Shani et al. (2020) make the first step to study the more difficult unknown transition and bandit feedback setting and propose an algorithm that achieves an $\widetilde{\mathcal{O}}(H^2 S\sqrt{A}K^{2/3})$ regret. The subsequent work by Luo et al. (2021) improves the result to $\widetilde{\mathcal{O}}(H^2 S\sqrt{AK})$. The second category solves adversarial MDPs using occupancy-measure-based algorithms. For the known transition setting, Zimin and Neu (2013) propose the O-REPS algorithm that achieves near-optimal regret for full-information and bandit feedback respectively. Rosenberg and Mansour (2019a) investigate the unknown transition but full-information setting. When the transition is unknown and only bandit feedback is available, Rosenberg and Mansour (2019b) propose an algorithm and prove it enjoys an $\widetilde{\mathcal{O}}(HS\sqrt{AK}/\alpha)$ regret with an addition assumption that all states are reachable with probability $\alpha > 0$ for any policy. Without this assumption, the regret bound degenerates to $\widetilde{\mathcal{O}}(H^{3/2}SA^{1/4}K^{3/4})$. Later, Jin et al. (2020a) achieve $\widetilde{\mathcal{O}}(H\sqrt{SAK})$ regret without the assumption of Rosenberg and Mansour (2019b). Finally, we remark that the existing tightest lower bound of $\Omega(H\sqrt{SAK})$ is established by Jin et al. (2018) for the unknown transition and full-information feedback, which also serves as a lower bound for the bandit feedback directly. In this work, we study the most challenging setting where the transition is unknown and only bandit feedback is available. Moreover, our solution falls into the second category, i.e., the occupancy-measure-based method.

RL with linear function approximation. To permit RL algorithm handling MDPs with large state and action space, a large body of literature considers solving MDPs with linear function approximation. In general, these studies can be categorized into three lines based on the specific assumption of the underlying MDP. The first line of work is according to the low Bellman-rank assumption (Jiang et al., 2017; Du et al., 2019), which assumes a low-rank factorization of the Bellman error matrix. The second line of work focuses on the linear MDPs (Yang and Wang, 2019; Jin et al., 2020b), where the transition kernel and loss function are parameterized as a linear function of a feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. The last line of work considers linear mixture/kernel MDPs (Ayoub et al., 2020; Zhou et al., 2021; Zhao et al., 2023b), where the transition kernel can be parameterized as a linear function of a feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$. Note that all the above works focus on MDPs with with linear function approximation under the *stochastic* loss functions. In this work, we investigate linear mixture MDPs but with the *adversarial* loss functions.

RL with adversarial losses and linear function approximation. Recent advances have emerged in learning adversarial RL with linear function approximation (Neu and Olkhovskaya, 2021; Zhong and Zhang, 2023; Sherman et al., 2023a; Luo et al., 2021; Dai et al., 2023; Sherman et al., 2023b; Kong et al., 2023; Liu et al., 2024; Cai et al., 2020; He et al., 2022; Li et al., 2023, 2024; Zhao et al., 2023a). Generally, these studies can be divided into two lines. The first line focuses on the linear MDPs. Neu and Olkhovskaya (2021) first study adversarial linear MDPs with bandit feedback but under the known transition setting. Zhong and Zhang (2023) first investigate the full-information and unknown transition setting and this setting is further studied by Sherman et al. (2023a) recently. Luo et al. (2021) make the first step to establish a sublinear regret for the more difficult unknown transition and bandit feedback setting. The result is further improved in (Dai et al., 2023; Sherman et al., 2023b; Kong et al., 2023; Liu et al., 2024). The second line of work considers the linear mixture MDPs. The seminal work of Cai et al. (2020) first studies adversarial linear mixture MDPs in the unknown transition and full-information feedback setting and proposes an optimistic proximal policy optimization algorithm. The subsequent work by He et al. (2022) improves their results to minimax optimality by using a weighted ridge regression and a Bernstein-type exploration bonus. The most recent work of Ji et al. (2024) studies the same setting and obtains a horizon-free regret which is independent of H with the assumption that the losses are upper bounded by $1/H$. For the more challenging unknown transition and bandit feedback setting, the only existing work of Zhao et al. (2023a) achieves a regret of $\widetilde{\mathcal{O}}(dS^2\sqrt{K} + \sqrt{HSAK})$, which exhibits a gap compared to the $\Omega(dH\sqrt{K} + \sqrt{HSAK})$ lower bound established in their work. In our work, we consider the same unknown transition and bandit feedback setting as Zhao et al. (2023a) and improve the upper bound to $\widetilde{\mathcal{O}}(d\sqrt{HS^3K} + \sqrt{HSAK})$, making a step towards closing the gap between the upper and lower bounds.

3 PROBLEM SETUP

In this section, we present the problem setup of episodic linear mixture MDPs with adversarial losses.

Episodic adversarial MDPs. In this paper, we consider episodic adversarial MDP, which is denoted by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{\ell_k\}_{k=1}^K)$. Here $\mathcal{S}$ is the state space with cardinality $|\mathcal{S}| = S$, $\mathcal{A}$ is the action space with cardinality $|\mathcal{A}| = A$, H is the length of each episode, K is the number of episodes, $P_h : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition kernel with $P_h(s' \mid s, a)$ is being the probability of transiting to state s' from state s and taking action a at stage h, $\ell_k : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the loss function, which may be chosen in an adversarial manner. Following previous studies (Zimin and Neu, 2013; Rosenberg and Mansour, 2019a; Jin et al., 2020a), we assume the MDP has a layered structure, satisfying the conditions:

- The state space $\mathcal{S}$ consists of $H + 1$ disjoint layers such that $\mathcal{S} = \cup_{h=1}^{H+1} \mathcal{S}_h$ and $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$ for $i \neq j$.
- $\mathcal{S}_1 = \{s_1\}$ and $\mathcal{S}_{H+1} = \{s_{H+1}\}$ are singletons.
- Transition is possible only between adjacent layers, that is $P_h(s' \mid s, a) = 0$ for all $s \in \mathcal{S}_h$ and $s' \notin \mathcal{S}_{h+1}$.

A policy $\pi = \{\pi_h\}_{h=1}^H$ is a collection of mapping π_h, where each $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ is a function maps a state s to distributions over $\mathcal{A}$ at stage h. Define the expected loss of an policy π at episode k as

$$L_k(\pi) = \mathbb{E}\left[\sum_{h=1}^{H} \ell_{k,h}(s_h, a_h) \;\middle|\; P, \pi\right], \tag{1}$$

where the expectation is taken over the randomness of the stochastic transition and policy.

In the *online MDP* setting, the interaction protocol between the learner and the environment is given as follows. The interaction proceeds in K episodes. At the beginning of episode k, the environment chooses a loss function ℓ_k, which may be in an adversarial manner. Simultaneously, the learner chooses a policy $\pi_k = \{\pi_{k,h}\}_{h=1}^H$. At each stage $h \in [H]$, the learner observes the state $s_{k,h}$, chooses an action $a_{k,h}$ sampled from $\pi_{k,h}(\cdot \mid s_{k,h})$, obtains reward $\ell_{k,h}(s_{k,h}, a_{k,h})$ and transits to the next state $s_{k,h+1} \sim P_h(\cdot \mid s_{k,h}, a_{k,h})$. In this work, we consider the *bandit feedback* setting where the learner can only observe the losses for the visited state-action pairs: $\{\ell_k(s_{k,h}, a_{k,h})\}_{h=1}^H$. The goal of the learner is to minimize regret, defined as

$$\mathrm{Reg}(K) = \sum_{k=1}^{K} L_k(\pi_k) - \sum_{k=1}^{K} L_k(\pi^*), \tag{2}$$

where $\pi^* \in \arg\min_{\pi \in \Pi} \sum_{k=1}^K L_k(\pi)$ is the optimal policy and Π is the set of all stochastic policy.

Linear Mixture MDPs. We focus on a special class of MDPs named *linear mixture MDPs* (Ayoub et al., 2020; Cai et al., 2020; Zhou et al., 2021; He et al., 2022; Li et al., 2023), where the transition kernel is linear in a known feature mapping $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ with the following definition.

Definition 1 (Linear Mixture MDPs)**.** An MDP instance $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, \{\ell_k\}_{k=1}^K)$ is called an inhomogeneous, episodic B-bounded linear mixture MDP if there exist a *known* feature mapping $\phi(s' \mid s, a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ with $\|\phi(s' \mid s, a)\|_2 \leq 1$ and *unknown* vectors $\{\theta_h^*\}_{h=1}^H \in \mathbb{R}^d$ with $\|\theta_h^*\|_2 \leq B$, such that for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and $h \in [H]$, it holds that $P_h(s' \mid s, a) = \langle \phi(s' \mid s, a), \theta_h^* \rangle$.

Occupancy measure. Previous studies (Zimin and Neu, 2013; Jin et al., 2020a) have shown the importance of the concept of *occupancy measure* for solving adversarial MDPs via online learning techniques. Specifically, for some policy π and transition kernel P, the occupancy measure $q^{P,\pi}$ is defined as the probability of visiting the state-action pair (s, a) when executing policy π under the transition P, that is

$$q^{P,\pi}(s, a) = \Pr[(s_h, a_h) = (s, a) \mid P, \pi], \tag{3}$$

where $h = h(s)$ is the index of the layer that state s belongs to. A valid occupancy measure q satisfies the following two properties. First, according to the loop-free structure, each layer is visited once and only once, and thus for all $h \in [H]$, we have $\sum_{s \in \mathcal{S}_h} \sum_{a \in \mathcal{A}} q(s, a) = 1$. Second, the probability of entering a state when coming from a previous layer is equal to the probability of leaving the state when going to the next layer, that is for all $h = 2, \ldots, H$ and $s \in \mathcal{S}_h$, we have $\sum_{(s',a') \in \mathcal{S}_{h-1} \times \mathcal{A}} q(s', a') P_{h-1}(s|s', a') = \sum_{a \in \mathcal{A}} q(s, a)$. Clearly, a valid occupancy measure q induce a policy π such that $\pi^q(a|s) = q(s, a) / \sum_{a' \in \mathcal{A}} q(s, a')$. For a fixed transition kernel P, we denote by $\Delta(P)$ the set of all valid occupancy measures induced by P. Similarly, we denote by $\Delta(\mathcal{P})$ the set of occupancy measures whose induced transition belongs to a set of transitions $\mathcal{P}$.

With the concept of occupancy measure, we can reduce this problem to the online linear optimization. Specifically, the expected loss of a policy π at episode k defined in (1) can be rewritten as

$$L_k(\pi) = \sum_{h=1}^{H} \sum_{s \in \mathcal{S}_h} \sum_{a \in \mathcal{A}} q^{P,\pi}(s, a) \ell_k(s, a) = \langle q^{P,\pi}, \ell_k \rangle.$$

Then the regret in (2) can be rewritten as

$$\mathrm{Reg}(K) = \sum_{k=1}^{K} \langle q^{P,\pi_k} - q^{P,\pi^*}, \ell_k \rangle. \tag{4}$$

We define $q^* \triangleq q^{P,\pi^*} \in \Delta(P)$ to simplify the notation.

4 THE PROPOSED ALGORITHM

This section introduces our proposed VLSUOB-REPS algorithm (Vector Least Square Upper Occupancy Bound Relative Entropy Policy Search) for adversarial linear mixture MDPs with unknown transition in the bandit feedback setting. VLSUOB-REPS consists of three key components: (i) estimating the unknown transition parameter and maintaining corresponding confidence set; (ii) constructing loss estimators; and (iii) applying online mirror descent over the occupancy measure space. We introduce the details below.

4.1 Transition Estimator

One of the main difficulties comes from the unknown transition kernel P. To address this issue, most existing works (Ayoub et al., 2020; Cai et al., 2020; Zhou et al., 2021) use the method of *value-targeted regression* (VTR) to learn the unknown parameter θ_h^* together with the corresponding confidence set. Specifically, for any function $V : \mathcal{S} \to \mathbb{R}$, define $\phi_V(s_{k,h}, a_{k,h}) = \sum_{s'} \phi(s' \mid s_{k,h}, a_{k,h}) V(s')$. By the definition of linear mixture MDPs in Definition 1, we have

$$P_h(\cdot \mid s_{k,h}, a_{k,h})^\top V(\cdot) = \langle \phi_V(s_{k,h}, a_{k,h}), \theta_h^* \rangle.$$

Therefore, learning the underlying θ_h^* can be regarded as solving a "linear bandit" problem (Lattimore and Szepesvári, 2020), where the context is $\phi_V(s_{k,h}, a_{k,h})$, and the noise is $V(s_{k,h+1}) - P_h(\cdot \mid s_{k,h}, a_{k,h})^\top V(\cdot)$. Thus, previous works (Ayoub et al., 2020; Cai et al., 2020) set the estimator $\theta_{k,h}$ as the minimizer of the least squares linear regression objective:

$$\sum_{i=1}^{k-1} \left[\phi_{V_{i,h+1}}(s_{i,h}, a_{i,h})^\top \theta - V_{i,h+1}(s_{i,h+1})\right]^2 + \lambda_k \|\theta\|_2^2,$$

where $V_{k,h}$ is the state value function defined as $V_{k,h}(s) = \mathbb{E}[\sum_{h'=h}^{H} \ell_{k,h}(s_{k,h}, a_{k,h}) \mid P, \pi_k, s_{k,h} = s]$. A similar *weighted* least squares linear regression method is used in the works (Zhou et al., 2021; He et al., 2022), which further utilizes the variance information of the value functions to gain a sharper confidence set.

Although value-targeted regression is the most popular method to estimate the unknown transition parameter in the literature, it is *not* applicable in our setting. The reason is that this method can only guarantee $\widehat{P}_h(\cdot \mid s, a)^\top V_{k,h+1}(\cdot) \approx P_h(\cdot \mid s, a)^\top V_{k,h+1}(\cdot)$, where $\widehat{P}$ is the estimated transition kernel. This method learns the transition kernel implicitly and bypasses the need for fully estimating the transition, which can be viewed as "model-free" in this sense. However, it is not sufficient for our purpose since the occupancy measure also depends on the transition kernel, which requires us to learn the transition kernel explicitly and ensure the estimated transition is accurate enough, i.e., $\widehat{P} \approx P$.

To this end, an alternative way to learn the unknown transition parameter is using the vanilla transition information directly. Specifically, denote $\Phi_{s,a} \in \mathbb{R}^{d \times S}$ with $\Phi_{s,a}(:, s') = \phi(s' \mid s, a)$ and let $\delta_s \in \{0, 1\}^S$ be the one-hot vector with $\delta_s(s) = 1$. Then, we can rewrite the transition kernel as $P_h(\cdot \mid s, a) = \Phi_{s,a}^\top \theta_h^*$. Thus, to learn the unknown parameter θ_h^*, we consider using $\Phi_{s_{k,h}, a_{k,h}}$ as feature and $\delta_{s_{k,h+1}}$ as the regression target. Then, the estimator $\theta_{k,h}$ is defined as the solution of the following linear regression problem:

$$\theta_{k,h} = \mathop{\arg\min}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{k-1} \left\| \Phi_{s_{i,h}, a_{i,h}}^\top \theta - \delta_{s_{i,h+1}} \right\|_2^2 + \lambda_k \|\theta\|_2^2. \tag{5}$$

The closed-form solution is $\theta_{k,h} = \Lambda_{k,h}^{-1} b_{k,h}$ with

$$\Lambda_{k,h} = \sum_{i=1}^{k-1} \sum_{s' \in \mathcal{S}_{h+1}} \phi(s' | s_{i,h}, a_{i,h}) \phi(s' | s_{i,h}, a_{i,h})^\top + \lambda_k I_d,$$

$$b_{k,h} = \sum_{i=1}^{k-1} \sum_{s' \in \mathcal{S}_{h+1}} \delta_{s_{i,h+1}}(s') \phi(s' \mid s_{i,h}, a_{i,h}). \tag{6}$$

Nonetheless, a significant challenge remains to be solved. Specifically, let $\varepsilon_{i,h} = P_h(\cdot \mid s_{i,h}, a_{i,h}) - \delta_{s_{i,h+1}}$ be the noise at episode i at stage h due to the transition. It is clear that $\varepsilon_{i,h} \in [-1, 1]^S$, $\mathbb{E}_{i,h}[\varepsilon_{i,h}] = \mathbf{0}$. One may consider establishing an ellipsoid confidence set for θ_h^* by applying the self-normalized concentration for vector-valued martingales (Abbasi-Yadkori et al., 2011, Theorem 1), as restated in Lemma 15 of Appendix E. However, since the learner only transits to one state in each layer, the noises across different states are *no longer independent*. Concretely, it hold that $\sum_{s \in \mathcal{S}} \varepsilon_{i,h}(s) = 0$. Thus the noises $\varepsilon_{i,h}(s)$ of different states are 1-subgaussian but they are not independent. This fact makes the key self-normalized concentration in Lemma 15 no longer applicable.

To address this challenge, Zhao et al. (2023a) propose to use the transition information of *only one* certain state $s'_{i,h+1}$ in the next layer, which they call the *imaginary* next state. They set the estimator $\theta_{k,h}$ as the minimizer of the following linear regression problem:

$$\sum_{i=1}^{k-1} \left[\phi(s'_{i,h+1} | s_{i,h}, a_{i,h})^\top \theta - \delta_{s_{i,h+1}}(s'_{i,h+1})\right]^2 + \lambda_k \|\theta\|_2^2. \tag{7}$$

Note that the imaginary next state $s'_{i,h+1}$ is not the actual next state $s_{i,h+1}$ experienced by the learner. Instead, they choose the imaginary next state $s'_{i,h+1}$

as the state with the largest uncertainty, formally,

$$s'_{i,h+1} = \arg\max_{s\in\mathcal{S}} \|\phi(s \mid s_{i,h}, a_{i,h})\|_{M_{i,h}^{-1}},$$

where $M_{i,h}$ is the feature covariance matrix, set as

$$\sum_{j=1}^{i-1} \phi(s'_{j,h+1} \mid s_{j,h}, a_{j,h})\phi(s'_{j,h+1} \mid s_{j,h}, a_{j,h})^\top + \lambda_i I.$$

With this choice, they can control the uncertainty of other states by that of the imaginary next state.

Though the method of using one state at each stage in Zhao et al. (2023a) is novel and provides an initial solution for this problem, it discards the visit information of other states and leads to a notable gap to the lower bound. To fully utilize the visit information, we use the information of *all* states and construct the estimator as in (5), instead of the only one state in (7). To address the non-independent noise issue, we introduce a new self-normalized concentration lemma tailored specifically for non-independent random noises. This lemma was originally proposed by Périvier and Goyal (2022) for the *dynamic assortment* problem, where a seller selects the subset of products to present to the customer who will then purchase one *single* item. They also face the non-independent random noises issue as the customer will only purchase at most *one* product, which is similar to our problem where the learner will only visit *one* state. Thus, they use this lemma to manage the product correlations and we make adaptations to handle the state correlations. Differently, Périvier and Goyal (2022, Theorem C.6) establish a variance-aware concentration inequality. In our work, we adapt their inequality into a simplified variance-independent form, which is well-suited for our analytical needs.

Lemma 1. *Let $\{\mathcal{F}_t\}_{t=0}^\infty$ be a filtration. Let $\{\delta_t\}_{t=1}^\infty$ be an $\mathbb{R}^N$-valued stochastic process such that δ_t is $\mathcal{F}_t$-measurable one-hot vector. Furthermore, assume $\mathbb{E}[\delta_t|\mathcal{F}_{t-1}] = p_t$ and define $\varepsilon_t = p_t - \delta_t$. Let $\{x_t\}_{t=1}^\infty$ be a sequence of $\mathbb{R}^{N\times d}$-valued stochastic process such that x_t is $\mathcal{F}_{t-1}$-measurable and $\|x_{t,i}\|_2 \le 1, \forall i \in [N]$. Let $\{\lambda_t\}_{t=1}^\infty$ be a sequence of non-negative scalars. Define*

$$Y_t = \sum_{i=1}^{t}\sum_{j=1}^{N} x_{i,j}x_{i,j}^\top + \lambda_t I_d, \quad S_t = \sum_{i=1}^{t}\sum_{j=1}^{N} \varepsilon_{i,j}x_{i,j}.$$

Then, for any $\zeta \in (0,1)$, with probability at least $1-\zeta$, we have for all $t \ge 1$,

$$\|S_t\|_{Y_t^{-1}} \le \frac{\sqrt{\lambda_t}}{4} + \frac{4}{\sqrt{\lambda_t}} \log\left(\frac{2^d \det(Y_t)^{\frac{1}{2}} \lambda_t^{-\frac{d}{2}}}{\zeta}\right).$$

With the above lemma, we can build the confidence set for the unknown parameter θ_h^* as follows.

Lemma 2. *Let $\zeta \in (0,1)$, then for any $k \in [K]$ and simultaneously for all $h \in [H]$, with probability at least $1-\zeta$, it holds that*

$$\theta_h^* \in \mathcal{C}_{k,h} \text{ where } \mathcal{C}_{k,h} = \{\theta \in \mathbb{R}^d \mid \|\theta - \theta_{k,h}\|_{\Lambda_{k,h}} \le \beta_k\}$$

with $\beta_k = (B+\frac{1}{4})\sqrt{\lambda_k} + \frac{2}{\sqrt{\lambda_k}}(2\log(\frac{H}{\zeta}) + d\log(4+\frac{4Sk}{\lambda_k d}))$.

Remark 1. Compared with the confidence set of $\|\theta - \theta_{k,h}\|_{M_{k,h}} \le \beta_k$ of Zhao et al. (2023a), our confidence set in Lemma 2 is tighter since $M_{k,h} \preceq \Lambda_{k,h}$. A primary challenge in constructing such a confidence set is bounding the self-normalized concentration term $\|\sum_{i=1}^k \sum_{s'\in\mathcal{S}_{h+1}} \varepsilon_{i,h+1}(s')\phi(s' \mid s_{i,h}, a_{i,h})\|_{\Lambda_{k,h}^{-1}}$. Due to the non-independent noises, we can not apply the self-normalized concentration in Lemma 15 directly. As a solution, Zhao et al. (2023a) propose to concentrate on a singular state per layer, which only need to bound the term $\|\sum_{i=1}^k \varepsilon_{i,h+1}(s'_{i,h+1})\phi(s'_{i,h+1} \mid s_{i,h}, a_{i,h})\|_{M_{k,h}^{-1}}$. While this approach bypasses the complications introduced by non-independent noises, it discards the visit information of other states. In contrast, we bound this challenging term by Lemma 1. This allows us to utilize the information of all states, leading to an improved bound. Intuitively, this new concentration lemma empowers our algorithm to explore the orientations of every state simultaneously, as opposed to the singular direction approach in Zhao et al. (2023a). ◁

Based on the above lemma, we can construct the confidence set $\mathcal{P}_k = \{\mathcal{P}_{k,h}\}_{h=1}^H$ for the transition P as

$$\mathcal{P}_{k,h} = \{\widehat{P}_h \mid \exists\theta \in \mathcal{C}_{k,h}, \widehat{P}_h(s'|s,a) = \phi(s'|s,a)^\top\theta\} \quad (8)$$

for all $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$. According to Lemma 2, we have $P \in \mathcal{P}_k$ with probability at least $1-\zeta$.

4.2 Loss Estimator

A common technique to deal with the bandit-feedback setting is to construct a loss estimator $\widehat{\ell}_{k,h}$ for the true loss function $\ell_{k,h}$ based on historical observations. When the transition is known, existing works (Zimin and Neu, 2013; Rosenberg and Mansour, 2019b) construct the unbiased estimator as

$$\widehat{\ell}_k(s,a) = \frac{\ell_k(s,a)}{q^{P,\pi_k}(s,a)} \mathbb{I}_k(s,a), \quad (9)$$

where $\mathbb{I}_k(s,a) = 1$ if (s,a) is visited at episode k and $\mathbb{I}_k(s,a) = 0$ otherwise. However, this method can not be directly applied to the unknown transition setting since the occupancy measure q^{P,π_k} is unknown. Rosenberg and Mansour (2019b) directly use the empirical occupancy measure $\widehat{q}^{P,\pi_k}$ in place of q^{P,π_k} to construct the estimator that could be either an overestimate or an underestimate, leading a loose regret bound.

To address this issue, Jin et al. (2020a) follow the principle of "optimistic in the face of uncertainty" and builds an underestimate for the loss function $\ell_{k,h}$ to encourage exploration. Since the true transition P belongs to the confidence set $\mathcal{P}_k$ with high probability. To build an underestimate for $\ell_{k,h}$, they propose to replace $q^{P,\pi_k}(s,a)$ in (9) with its *upper occupancy bound*, defined as the largest possible probability of visiting (s,a) under the confidence set $\mathcal{P}_k$. Formally,

$$u_k(s,a) = \max_{\widehat{P} \in \mathcal{P}_k} q^{\widehat{P},\pi_k}(s,a). \tag{10}$$

The above step can be computed efficiently by the Comp-UOB procedure of Jin et al. (2020a). Additionally, they adopt the idea of *implicit exploration* of Neu (2015) to further increase the denominator by some fixed amount $\gamma > 0$, which is for several technical reasons such as obtaining a high probability bound. Finally, the estimator is built as

$$\widehat{\ell}_k(s,a) = \frac{\ell_k(s,a)}{u_k(s,a) + \gamma} \mathbb{I}_k(s,a). \tag{11}$$

Clearly, $\widehat{\ell}_k(s,a)$ is an underestimate of $\ell_k(s,a)$ since $u_k(s,a) \geq q^{P,\pi_k}(s,a)$ with high probability.

In our algorithm, we follow the work of Jin et al. (2020a) and employ the loss estimator defined in (11).

4.3 Online Mirror Descent

Online Mirror Descent (OMD) is a powerful framework for solving online convex optimization problems (Orabona, 2019; Wei and Luo, 2018; Zhao et al., 2021). As discussed in Section 3, our problem is closely related to the online linear optimization problem over the occupancy measure space. Thus, we utilize OMD as a key component of our algorithm. We apply OMD over the occupancy measure space $\Delta(\mathcal{P}_k)$ induced by the confidence set $\mathcal{P}_k$. Specifically, we update the occupancy measure as follows:

$$\widehat{q}_{k+1} = \mathop{\arg\min}_{q \in \Delta(\mathcal{P}_k)} \eta \langle q, \widehat{\ell}_k \rangle + D_\psi(q \| \widehat{q}_k), \tag{12}$$

where $\widehat{\ell}_k$ is the loss estimator defined in (11), $\eta > 0$ is step size, $\psi(q) = \sum_{s,a} q(s,a) \log q(s,a) - \sum_{s,a} q(s,a)$ is the unnormalized negative entropy, and $D_\psi(q\|q') = \sum_{s,a} q(s,a) \log \frac{q(s,a)}{q'(s,a)} - \sum_{s,a}(q(s,a) - q'(s,a))$ is the unnormalized KL-divergence. The update procedure in (12) can also be implemented efficiently, as discussed in Appendix E of Zhao et al. (2023a).

The detailed algorithm is presented in Algorithm 1. Line 3 - 9 estimate the unknown transition, Line 10 compute the upper occupancy bound u_k, Line 11 constructs the loss estimator $\widehat{\ell}_k$, and Line 12 runs OMD to update the occupancy measure $\widehat{q}_{k+1}$. The learner execute the policy π_{k+1} induced by $\widehat{q}_{k+1}$ in Line 13.

Algorithm 1 VLSUOB-REPS

Input: Confidence parameter ζ, step size η, regularization parameter λ_k, exploration parameter γ.

1: **Initialization:** Set confidence set $\mathcal{P}_1$ as all transition kernels. For all $h \in [H]$ and all $s \in \mathcal{S}_h$, set $\widehat{q}_1(s,a) = \frac{1}{S_h \times A}$. Let $\pi_1 = \pi^{\widehat{q}_1}$, $\Lambda_{1,h} = \lambda_1 I_d, \forall h$.
2: **for** $k = 1, \ldots, K$ **do**
3:   **for** $h = 1, \ldots, H$ **do**
4:     Take action $a_{k,h} \sim \pi_{k,h}(\cdot \mid s_{k,h})$.
5:     Suffer and observe loss $\ell_{k,h}(s_{k,h}, a_{k,h})$.
6:     Transit to $s_{k,h+1} \sim P_h(\cdot \mid s_{k,h}, a_{k,h})$.
7:     $\theta_{k,h} = \Lambda_{k,h}^{-1} b_{k,h}$ with $\Lambda_{k,h}$ and $b_{k,h}$ as in (6).
8:     Construct the confidence set as in (8).
9:   **end for**
10:   Compute upper bound $u_k(s,a)$ as in (10).
11:   Construct loss estimator $\widehat{\ell}_{k,h}$ as in (11).
12:   Compute occupancy measure $\widehat{q}_{k+1}$ as in (12).
13:   Update policy $\pi_{k+1} = \pi^{\widehat{q}_{k+1}}$.
14: **end for**

5 REGRET GUARANTEE

In this section, we present the regret upper bound of our algorithm and the proof sketch.

5.1 Regret Upper Bound

The regret bound of our algorithm VLSUOB-REPS is guaranteed by the following theorem.

Theorem 1. *Set the step size η and exploration parameter γ as $\eta = \gamma = \sqrt{\frac{H \log(HSA/\zeta)}{KSA}}$, the regularization parameter λ_k as $\lambda_1 = 1, \lambda_k = d \log(kS)$, $\forall k > 1$. With probability at least $1 - 5\zeta$, VLSUOB-REPS algorithm ensures the regret $\mathrm{Reg}(K)$ is upper bounded by*

$$\mathcal{O}\left(d\sqrt{HS^3K} \log^2\left(\frac{dSK}{\zeta}\right) + \sqrt{HSAK \log\left(\frac{HSA}{\zeta}\right)} \right).$$

Remark 2. Compared with the regret bound of $\widetilde{\mathcal{O}}(dS^2\sqrt{K} + \sqrt{HSAK})$ in Zhao et al. (2023a, Theorem 1), our bound is better since $H \leq S$ by the layered structure of MDPs. As a byproduct, our result improves the best-known $\widetilde{\mathcal{O}}(HS\sqrt{AK})$ regret for tabular MDPs (Jin et al., 2020a) when $d \leq \sqrt{HA/S}$. ◁

Remark 3. Compared with the lower bound of $\Omega(dH\sqrt{K} + \sqrt{HSAK})$ established in Zhao et al. (2023a, Theorem 2), our regret is suboptimal in the dependence on S. However, note that the dependence on S remains suboptimal even for tabular MDPs (Jin et al., 2020a). How to close this gap is an important open question and we leave it as future work. ◁

5.2 Occupancy Measure Difference

In this part, we introduce a key technical lemma that bounds the occupancy measure difference induced by the different transitions in the confidence set and is critical in our analysis.

Lemma 3 (Occupancy measure difference for linear mixture MDPs)**.** *For any collection of transition kernels $\{P_k^s\}_{s\in\mathcal{S}}$ such that $P_k^s \in \mathcal{P}_k$ for all $s \in \mathcal{S}$, if $\lambda \geq \delta$, with probability at least $1-2\zeta$, it holds that*

$$\sum_{k=1}^{K} \left\| q^{P_k^s,\pi_k} - q_k \right\|_1 \leq \mathcal{O}\left(d\sqrt{HS^3K}\log^2(dSK/\zeta)\right).$$

Remark 4. Compared with the occupancy measure difference $\widetilde{\mathcal{O}}(dS^2\sqrt{K})$ of Zhao et al. (2023a, Lemma 2), our bound $\widetilde{\mathcal{O}}(d\sqrt{HS^3K})$ in Lemma 3 is better since $H \leq S$ by the layed structure of MDPs. This improvement comes from the new self-normalized concentration for non-independent random noises in Lemma 1, which allows us to use the transition information of all states instead of only one as in Zhao et al. (2023a). ◁

Remark 5. Compared with the occupancy measure difference $\widetilde{\mathcal{O}}(HS\sqrt{AK})$ of Jin et al. (2020a, Lemma 4), our bound $\widetilde{\mathcal{O}}(d\sqrt{HS^3K})$ in Lemma 3 is better when $d \leq \sqrt{HA/S}$. Though our bound gets rid of the dependence on the action space size A, it still keeps the dependence on the state space S. As pointed by Zhao et al. (2023a), the main hardness of simultaneously eliminating the dependence of the occupancy measure difference on both S and A is that though the transition kernel P of a linear mixture MDP admits a linear structure, the occupancy measure still has a complicated recursive form: $q_k(s,a) = \pi_k(a|s)\langle\theta^*_{h(s)-1}, \sum_{(s',a')\in\mathcal{S}_{h(s)-1}\times\mathcal{A}} q_k(s',a')\,\phi(s|s',a')\rangle$. We leave the question of whether it is possible to eliminate the dependence on S as future work. ◁

Next, we present the proof sketch of Lemma 3.

Proof Sketch (of Lemma 3)**.** First, we introduce the following lemma, which bounds the error between the transition P_h and the estimated transition $\widehat{P}_{k,h} \in \mathcal{P}_{k,h}$.

Lemma 4. *Let $\widehat{P}_k = \{\widehat{P}_{k,h}\}_{h=1}^H$ with $\widehat{P}_{k,h} \in \mathcal{P}_{k,h}$ such that $\widehat{P}_{k,h}(s' \mid s,a) = \phi(s' \mid s,a)^\top \widehat{\theta}_{k,h}$ for all $(s,a,s') \in \mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}$ for some $\widehat{\theta}_{k,h} \in \mathcal{C}_{k,h}$. Then for any $\zeta \in (0,1)$ and simultaneously for all $k \in [K]$ and $h \in [H]$, with probability at least $1-\zeta$, it holds that*

$$|\widehat{P}_{k,h}(s'|s,a) - P_h(s'|s,a)| \leq 1 \wedge \beta_k \|\phi(s' \mid s,a)\|_{\Lambda_{k,h}^{-1}},$$

where β_k is the diameter of confidence set defined in Lemma 2 and $\Lambda_{k,h}$ is covariance matrix defined in (6).

Then, we present the following lemma, which bounds the error between the occupancy measure by the estimation error of the transition. For the sake of brevity, we define $\epsilon_{k,h}(s' \mid s,a) = 1 \wedge \beta_k\|\phi(s' \mid s,a)\|_{\Lambda_{k,h}^{-1}}$, then we have $|\widehat{P}_{k,h}(s'|s,a) - P_h(s'|s,a)| \leq \epsilon_{k,h}(s' \mid s,a)$ by Lemma 4. Then we have the following lemma.

Lemma 5. *For any collection of transition kernels $\{P_k^s\}_{s\in\mathcal{S}}$ such that $P_k^s \in \mathcal{P}_k$ for all $s \in \mathcal{S}$, if $\lambda \geq \zeta$, with probability at least $1-2\zeta$, it holds that*

$$\sum_{k=1}^{K} \left\| q^{P_k^s,\pi_k} - q_k \right\|_1 \leq 2S \sum_{k,h,s'} \epsilon_{k,h}^{s'} + 4\beta_K S^2 \log\left(\frac{H}{\zeta}\right).$$

Here $\sum_{k,h,s'} \epsilon_{k,h}^{s'} \triangleq \sum_{k,h}\sum_{s'\in\mathcal{S}_{h+1}} \epsilon_{k,h}(s'|s_{k,h},a_{k,h})$.

According to Lemma 5, to bound the occupancy measure difference, it suffices to bound the estimation error of the transition, that is the cumulative error of $\epsilon_{k,h}(s' \mid s_{k,h},a_{k,h})$ over all episodes k and all stages h. Bounding this term is the main difference between our work and Zhao et al. (2023a). In particular, Zhao et al. (2023a) only use the transition information of one state $s'_{k,h+1}$ with the maximum uncertainty in the next layer. Thus they can bound the estimation error of other states by that of $s'_{k,h+1}$. Specifically, note that $\epsilon_{k,h}(s'|s,a) \leq \beta_k(1 \wedge \|\phi(s'|s,a)\|_{\Lambda_{k,h}^{-1}})$, they bound the cumulative error of transition as below:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}_{h+1}} \beta_k(1 \wedge \|\phi(s' \mid s_{k,h},a_{k,h})\|_{M_{k,h}^{-1}}) \\
&\leq \beta_K \sum_{k=1}^{K}\sum_{h=1}^{H} S_{h+1}(1 \wedge \|\phi(s'_{k,h+1} \mid s_{k,h},a_{k,h})\|_{M_{k,h}^{-1}}) \\
&\leq \beta_K S\sqrt{K\sum_{k=1}^{K}\left(1 \wedge \|\phi(s'_{k,h+1} \mid s_{k,h},a_{k,h})\|_{M_{k,h}^{-1}}\right)^2} \\
&\leq \widetilde{\mathcal{O}}\big(\beta_K S\sqrt{dK}\big).
\end{aligned}$$

Here the first inequality holds by the choice that $s'_{i,h+1} = \arg\max_{s\in\mathcal{S}}\|\phi(s \mid s_{i,h},a_{i,h})\|_{M_{i,h}^{-1}}$, the second inequality holds by the Cauchy-Schwarz inequality, the last inequality follows from the self-normalized concentration of Abbasi-Yadkori et al. (2011, Lemma 10).

Instead, we use the transition information of all states and thus can bound this term directly as follows

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}_{h+1}} \beta_k(1 \wedge \|\phi(s' \mid s_{k,h},a_{k,h})\|_{\Lambda_{k,h}^{-1}}) \\
&\leq \beta_K \sum_{h=1}^{H}\sqrt{KS_{h+1}\sum_{k=1}^{K}\sum_{s'\in\mathcal{S}_{h+1}}(1 \wedge \|\phi(s' \mid s_{k,h}a_{k,h})\|_{\Lambda_{k,h}^{-1}})^2} \\
&\leq \widetilde{\mathcal{O}}\big(\beta_K\sqrt{dHSK}\big),
\end{aligned}$$

which replaces a dependence on $\sqrt{S}$ with $\sqrt{H}$, resulting the final improvement over Zhao et al. (2023a). ■

5.3 Proof Sketch of Theorem 1

Finally, we present the proof sketch of Theorem 1.

Define the occupancy measure under the true transition P and policy π_k as $q_k = q^{P,\pi_k}$. Then, the regret can be written as $\text{Reg} = \sum_{k=1}^K \langle q_k - q^*, \ell_k \rangle$. Following the work of Jin et al. (2020a), we decompose the regret as the following four terms:

$$\text{Reg}(K) \leq \underbrace{\sum_{k=1}^K \langle \widehat{q}_k - q^*, \widehat{\ell}_k \rangle}_{\text{Regret}} + \underbrace{\sum_{k=1}^K \langle q_k - \widehat{q}_k, \ell_k \rangle}_{\text{Error}} + \underbrace{\sum_{k=1}^K \langle \widehat{q}_k, \ell_k - \widehat{\ell}_k \rangle}_{\text{Bias-I}} + \underbrace{\sum_{k=1}^K \langle q^*, \widehat{\ell}_k - \ell_k \rangle}_{\text{Bias-II}}.$$

Here, the first Regret term is the regret of the corresponding online linear optimization problem with respect to the loss estimator $\widehat{\ell}_k$, which can be controlled by OMD via standard analysis and can be bounded by $\mathcal{O}(\sqrt{HSAK \log(HSA/\zeta)} + H \log(H/\zeta))$. The last Bias-II term measures the bias of the loss estimator $\widehat{\ell}_k$ with respect to the true loss ℓ_k, which can be bounded by $\mathcal{O}(\sqrt{HSAK \log(SA/\zeta)})$ by the concentration of the implicit exploration (Neu, 2015). Finally, the remaining two terms Error and Bias-I come from the error of using $\widehat{q}_k$ and u_k to approximate q_k respectively, which are closely related to the occupancy measure difference in Lemma 3. We bound these two terms in the rest. Bounding Regret and Bias-II is relatively standard and we defer the proofs to Appendix D.

Bounding Error Term. With Lemma 3, we immediately obtain the following bound on Error.

Lemma 6. *For any $\zeta \in (0,1)$, with probability at least $1-2\zeta$, VLSUOB-REPS algorithm ensures that*

$$\text{Error} \leq \mathcal{O}\Big(d\sqrt{HS^3K} \log^2(dKS/\zeta)\Big).$$

Proof. Let $P_k^s = P^{\widehat{q}_k} \in \mathcal{P}_k$ for all s such that $\widehat{q}_k = q^{P_k,\pi_k}$. Since $\ell_k(s,a) \in [0,1]$ for all $k \in [K]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $\text{Error} \leq \sum_{k=1}^K \sum_{s,a} |\widehat{q}_k(s,a) - q_k(s,a)| = \sum_{k=1}^K \sum_{s,a} |q^{P_k^s,\pi_k} - q_k(s,a)|$. The proof is then completed by applying Lemma 3. ■

Bounding Bias-I Term. To bound this term, we need to show the loss estimator $\widehat{\ell}_k$ is close to the true loss function ℓ_k. This is guaranteed by the fact that the confidence set becomes more and more accurate for frequently visited state-action pairs.

Lemma 7. *For any $\zeta \in (0,1)$, with probability at least $1-3\zeta$, VLSUOB-REPS algorithm ensures that*

$$\text{Bias-I} \leq \mathcal{O}\left(d\sqrt{HS^3K} \log^2(dKS/\zeta) + \gamma SAK\right).$$

Proof. To bound the Bias-I, we first bound the term $\sum_{k=1}^K \langle \widehat{q}_k, \ell_k - \mathbb{E}_{k-1,H}[\widehat{\ell}_k] \rangle$ as follows.

$$\begin{aligned}
&\sum_k \left\langle \widehat{q}_k, \ell_k - \mathbb{E}_{k-1,H}\left[\widehat{\ell}_k\right] \right\rangle \\
&= \sum_{k,s,a} \widehat{q}_k(s,a)\ell_k(s,a)\left(1 - \frac{\mathbb{E}_{k-1,H}\left[\mathbb{I}_k\{s,a\}\right]}{u_k(s,a)+\gamma}\right) \\
&= \sum_{k,s,a} \widehat{q}_k(s,a)\ell_k(s,a)\left(1 - \frac{q_k(s,a)}{u_k(s,a)+\gamma}\right) \\
&\leq \sum_{k,s,a} |u_k(s,a) - q_k(s,a)| + \gamma SAK.
\end{aligned}$$

Since $u_k = q^{P_k^s,\pi_k}$ where $P_k^s = \arg\max_{\widehat{P}\in\mathcal{P}_k} q^{\widehat{P},\pi_k}(s)$, the term $\sum_{k,s,a} |u_k(s,a) - q_k(s,a)|$ can be controlled by Lemma 3 again. It remains to bound the term $\sum_{k=1}^K \langle \widehat{q}_k, \mathbb{E}_{k-1,H}[\widehat{\ell}_k] - \widehat{\ell}_k \rangle$. Since the fact that $P^{\widehat{q}_k} \in \mathcal{P}_k$ and $u_k(s,a) = \max_{\widehat{P}\in\mathcal{P}_k} q^{\widehat{P},\pi_k}(s,a)$, we have $\sum_{s,a} \widehat{q}_k(s,a)\widehat{\ell}_k(s,a) \leq \sum_{s,a} u_k(s,a)\widehat{\ell}_k(s,a) = H$. Then using the Azuma-Hoeffding inequality, with probability at least $1-\zeta$, we have

$$\sum_{k=1}^K \langle \widehat{q}_k, \mathbb{E}_{k-1,H}[\widehat{\ell}_k] - \widehat{\ell}_k \rangle \leq H\sqrt{2K \log(1/\zeta)}.$$

Applying the union bound finishes the proof. ■

6 CONCLUSION

In this work, we consider learning adversarial linear mixture MDPs with bandit feedback and unknown transition. We propose a new algorithm that achieves an $\widetilde{\mathcal{O}}(d\sqrt{HS^3K} + \sqrt{HSAK})$ regret with high probability. Our result strictly improves the previously best-known $\widetilde{\mathcal{O}}(dS^2\sqrt{K} + \sqrt{HSAK})$ regret (Zhao et al., 2023a). As a byproduct, it improves the best-known $\widetilde{\mathcal{O}}(HS\sqrt{AK})$ result for tabular MDPs (Jin et al., 2020a) when $d \leq \sqrt{HA/S}$. To achieve this result, we first propose a new least square estimator for the unknown transition that leverages the visit information of all states, as opposed to only a single state in Zhao et al. (2023a). Then we introduce a new self-normalized concentration designed specifically for non-independent noises to handle the state correlations.

Several questions remain open for future study. First, the dependence on S of our result is suboptimal, and how to close this gap is an important open problem. Moreover, optimizing the dynamic regret of adversarial MDPs is an emerging direction to facilitate algorithms with more robustness in non-stationary environments. Recent literature has explored the dynamic regret of adversarial MDPs with full-information feedback (Fei et al., 2020; Zhao et al., 2022; Li et al., 2023). Extending their results to the bandit feedback setting is an important future direction.

Acknowledgements

This research was supported by NSFC (62206125, U23A20382, 61921006). Peng Zhao was supported in part by the Xiaomi Foundation.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 2312–2320.

Ayoub, A., Jia, Z., Szepesvári, C., Wang, M., and Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 463–474.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 263–272.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. E. (2011). Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 19–26.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2020). Provably efficient exploration in policy optimization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1283–1294.

Dai, Y., Luo, H., Wei, C., and Zimmert, J. (2023). Refined regret for adversarial MDPs with linear function approximation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 6726–6759.

Du, S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudik, M., and Langford, J. (2019). Provably efficient RL with rich observations via latent state decoding. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 1665–1674.

Even-Dar, E., Kakade, S. M., and Mansour, Y. (2009). Online Markov decision processes. *Mathematics of Operations Research*, pages 726–736.

Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. (2020). Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3052–3060.

Fei, Y., Yang, Z., Wang, Z., and Xie, Q. (2020). Dynamic regret of policy optimization in non-stationary environments. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 6743–6754.

He, J., Zhou, D., and Gu, Q. (2022). Near-optimal policy optimization algorithms for learning adversarial linear mixture MDPs. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4259–4280.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, pages 1563–1600.

Ji, K., Zhao, Q., He, J., Zhang, W., and Gu, Q. (2024). Horizon-free reinforcement learning in adversarial linear mixture MDPs. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1704–1713.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 4868–4878.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. (2020a). Learning adversarial Markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 4860–4869.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020b). Provably efficient reinforcement learning with linear function approximation. In *Proceedings of the 33rd Conference on Learning Theory (COLT)*, pages 2137–2143.

Kong, F., Zhang, X., Wang, B., and Li, S. (2023). Improved regret bounds for linear adversarial MDPs via linear optimization. *Transactions on Machine Learning Research.*

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms.* Cambridge University Press.

Li, L.-F., Zhao, P., and Zhou, Z.-H. (2023). Dynamic regret of adversarial linear mixture MDPs. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 60685–60711.

Li, L.-F., Zhao, P., and Zhou, Z.-H. (2024). Dynamic regret of adversarial MDPs with unknown transition and linear function approximation. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*, page to appear.

Liu, H., Wei, C.-Y., and Zimmert, J. (2024). Towards optimal regret in adversarial linear MDPs with bandit feedback. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Luo, H., Wei, C., and Lee, C. (2021). Policy optimization in adversarial MDPs: Improved exploration via dilated bonuses. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 22931–22942.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. A. (2013). Playing Atari with deep reinforcement learning. *ArXiv preprint*, 1312.5602.

Neu, G. (2015). Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, pages 3168–3176.

Neu, G. and Olkhovskaya, J. (2021). Online learning in MDPs with linear function approximation and bandit feedback. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 10407–10417.

OpenAI (2023). GPT-4 technical report. *ArXiv preprint*, 2303.08774.

Orabona, F. (2019). A modern introduction to online learning. *ArXiv preprint*, 1912.13213.

Périvier, N. and Goyal, V. (2022). Dynamic pricing and assortment under a contextual MNL demand. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pages 3461–3474.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley.

Rosenberg, A. and Mansour, Y. (2019a). Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5478–5486.

Rosenberg, A. and Mansour, Y. (2019b). Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 2209–2218.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *ArXiv preprint*, 1707.06347.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. (2020). Optimistic policy optimization with bandit feedback. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 8604–8613.

Sherman, U., Cohen, A., Koren, T., and Mansour, Y. (2023a). Rate-optimal policy optimization for linear Markov decision processes. *ArXiv preprint*, arXiv:2308.14642.

Sherman, U., Koren, T., and Mansour, Y. (2023b). Improved regret for efficient online reinforcement learning with linear function approximation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 31117–31150.

Wei, C.-Y. and Luo, H. (2018). More adaptive algorithms for adversarial bandits. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, pages 1263–1291.

Yang, L. and Wang, M. (2019). Sample-optimal parametric Q-learning using linearly additive features. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6995–7004.

Yu, J. Y., Mannor, S., and Shimkin, N. (2009). Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, pages 737–757.

Zhao, C., Yang, R., Wang, B., and Li, S. (2023a). Learning adversarial linear mixture Markov decision processes with bandit feedback and unknown transition. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Zhao, H., He, J., Zhou, D., Zhang, T., and Gu, Q. (2023b). Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In *The 36th Annual Conference on Learning Theory (COLT)*, pages 4977–5020.

Zhao, P., Li, L.-F., and Zhou, Z.-H. (2022). Dynamic regret of online Markov decision processes. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 26865–26894.

Zhao, P., Zhang, Y.-J., Zhang, L., and Zhou, Z.-H. (2021). Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *ArXiv preprint*, 2112.14368.

Zhong, H. and Zhang, T. (2023). A theoretical analysis of optimistic proximal policy optimization in linear Markov decision processes. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pages 73666–73690.

Zhou, D., Gu, Q., and Szepesvári, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, pages 4532–4576.

Zimin, A. and Neu, G. (2013). Online learning in episodic Markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 1583–1591.

A Proof of Lemma 1

In this section, we present the proof of Lemma 1, which is a simplified version of the proof of Périvier and Goyal (2022, Theorem C.6). For self-containedness, we present the proof below.

A.1 Main Proof

Proof. We define a global variable as $z_i = \sum_{j=1}^{N} \varepsilon_{i,j} x_{i,j}$ and analyze the concentration of z_i. Denote $\mathcal{B}_d(x, r)$ the d-dimensional ℓ_2 ball centered at x with radius r. For all $\xi \in \mathcal{B}_d(0, 1/2)$, we define

$$M_0(\xi) = 1, \text{ and } M_t(\xi) = \exp(\xi^\top z_t - \|\xi\|_{Y_t}^2). \tag{13}$$

We denote by $\mathcal{F}_t$ the σ-algebra generated by $\{\{x_i, \varepsilon_i\}_{i=1}^{t-1}, x_t\}$. To prove Lemma 1, the crucial step is to demonstrate that though z_i is a combination of non-independent variables, relatively to $\mathcal{F}_t$, $M_t(\xi)$ is still a super martingale, which is present in the following lemma.

Lemma 8. *For all $\xi \in \mathcal{B}_d(0, 1/2)$, $\{M_t(\xi)\}_{t=0}^{\infty}$ defined in (13) is a non-negative super martingale.*

We present the proof of Lemma 8 in Appendix A.2. Then, the remaining proof follows the proof of Faury et al. (2020, Theorem 1). The main difference in our analysis is that ξ belongs to $\mathcal{B}_d(0, 1/2)$ instead of $\mathcal{B}_d(0, 1)$ to ensure $\{M_t(\xi)\}_{t=0}^{\infty}$ is a super martingale.

Let $h(\xi)$ be a probability density function with support on $\mathcal{B}_d(0, 1/2)$. For $t \geq 0$ let:

$$\bar{M}_t \triangleq \int_{\xi} M_t(\xi) dh(\xi)$$

By Lemma 20.3 of Lattimore and Szepesvári (2020), $\bar{M}_t$ is also a non-negative super-martingale, and $\mathbb{E}\left[\bar{M}_0\right] = 1$. Let τ be a stopping time with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. We can follow the proof of Lemma 8 in Abbasi-Yadkori et al. (2011) to justify that $\bar{M}_\tau$ is well-defined (independently of whether $\tau < \infty$ holds or not) and that $\mathbb{E}\left[\bar{M}_\tau\right] \leq 1$. Therefore, with $\zeta \in (0, 1)$ and thanks to the maximal inequality:

$$\Pr\left[\log\left(\bar{M}_\tau\right) \geq \log\left(\frac{1}{\zeta}\right)\right] = \Pr\left[\bar{M}_\tau \geq \frac{1}{\zeta}\right] \leq \zeta.$$

Then, we compute the lower bound of $\bar{M}_t$ as follows. Let $\beta_t = \sqrt{2\lambda_t}$ be a positive scalar and set h to be the density function of an isotropic normal distribution of precision β_t^2 truncated on $\mathcal{B}_d(0, 1/2)$. Denote $N(h)$ its normalization constant. Then, we have

$$\bar{M}_t = \frac{1}{N(h)} \int_{\mathcal{B}_d(0,1/2)} \exp\left(\xi^\top S_t - \|\xi\|_{Y_t}^2\right) d\xi.$$

To simplify the notation, let $f(\xi) := \xi^\top S_t - \|\xi\|_{Y_t}^2$ and $\xi_* = \arg\max_{\|\xi\|_2 \leq 1/4} f(\xi)$, we obtain

$$\begin{aligned}
\bar{M}_t &= \frac{e^{f(\xi_*)}}{N(h)} \int_{\mathbb{R}^d} \mathbf{1}_{\|\xi\|_2 \leq 1/2} \exp\Big((\xi - \xi_*)^\top \nabla f(\xi_*) - (\xi - \xi_*)^\top Y_t (\xi - \xi_*)\Big) d\xi \\
&= \frac{e^{f(\xi_*)}}{N(h)} \int_{\mathbb{R}^d} \mathbf{1}_{\|\xi + \xi_*\|_2 \leq 1/2} \exp\left(\xi^\top \nabla f(\xi_*) - \xi^\top Y_t \xi\right) d\xi \\
&\geq \frac{e^{f(\xi_*)}}{N(h)} \int_{\mathbb{R}^d} \mathbf{1}_{\|\xi\|_2 \leq 1/4} \exp\left(\xi^\top \nabla f(\xi_*) - \xi^\top Y_t \xi\right) d\xi \\
&= \frac{e^{f(\xi_*)}}{N(h)} \int_{\mathbb{R}^d} \mathbf{1}_{\|\xi\|_2 \leq 1/4} \exp\left(\xi^\top \nabla f(\xi_*)\right) \exp\left(-\frac{1}{2}\xi^\top (2Y_t)\, \xi\right) d\xi.
\end{aligned}$$

Further, we define $g(\xi)$ as the density of the normal distribution of precision $2Y_t$ truncated on the ball $\mathcal{B}_d(0, 1/4)$ and $N(g)$ its normalization constant. Then, we have

$$\bar{M}_t \geq \exp\left(f(\xi_*)\right) \frac{N(g)}{N(h)} \mathbb{E}_g\left[\exp\left(\xi^\top \nabla f(\xi_*)\right)\right] \geq \exp\left(f(\xi_*)\right) \frac{N(g)}{N(h)} \exp\left(\mathbb{E}_g\left[\xi^\top \nabla f(\xi_*)\right]\right) \geq \exp\left(f(\xi_*)\right) \frac{N(g)}{N(h)},$$

where the last inequality holds by $\mathbb{E}_g[\xi] = 0$. Then, we obtain that for all ξ_0 such that $\|\xi_0\|_2 \leq 1/4$:

$$\begin{aligned}
\Pr\left[\bar{M}_t \geq \frac{1}{\zeta}\right] &\geq \Pr\left[\exp\left(f\left(\xi_*\right)\right)\frac{N(g)}{N(h)} \geq 1/\zeta\right] \\
&= \Pr\left[\log\left(\exp\left(f\left(\xi_*\right)\right)\frac{N(g)}{N(h)}\right) \geq \log(1/\zeta)\right] \\
&= \Pr\left[f\left(\xi_*\right) \geq \log(1/\zeta) + \log\left(\frac{N(h)}{N(g)}\right)\right] \\
&= \Pr\left[\max_{\|\xi\|_2 \leq 1/4} \xi^\top S_t - \|\xi\|_{Y_t}^2 \geq \log(1/\zeta) + \log\left(\frac{N(h)}{N(g)}\right)\right] \\
&\geq \Pr\left[\xi_0^\top S_t - \|\xi_0\|_{Y_t}^2 \geq \log(1/\zeta) + \log\left(\frac{N(h)}{N(g)}\right)\right].
\end{aligned}$$

In particular, we set $\xi_0 = \frac{Y_t^{-1} S_t}{\|S_t\|_{Y_t^{-1}}} \frac{\beta_t}{4\sqrt{2}} \leq \frac{1}{4}$. Thus, we have

$$\Pr\left[\|S_t\|_{Y_t^{-1}} \geq \frac{\beta_t}{4\sqrt{2}} + \frac{4\sqrt{2}}{\beta_t}\log\left(\frac{N(h)}{\zeta N(g)}\right)\right] \leq \Pr\left[\bar{M}_t \geq \frac{1}{\zeta}\right]. \tag{14}$$

It remains to bound the quantities $N(h)$ and $g(h)$. To this end, we introduce the following lemma in Faury et al. (2020) which bounds the log of their ratio.

Lemma 9 (Lemma 6 of Faury et al. (2020))**.** *The following inequality holds:*

$$\log\left(\frac{N(h)}{N(g)}\right) \leq \log\left(\frac{2^{d/2}\det\left(Y_t\right)^{1/2}}{\beta_t^d}\right) + d\log(2). \tag{15}$$

Combining (14) and (15), with probability at least $1-\zeta$, for all t it holds that

$$\|S_t\|_{Y_t^{-1}} \leq \frac{\beta_t}{4\sqrt{2}} + \frac{4\sqrt{2}}{\beta_t}\log\left(\frac{2^{d/2}\det\left(Y_t\right)^{1/2}}{\beta_t^d \zeta}\right) + \frac{4\sqrt{2}}{\beta_t} d\log(2).$$

Finally, the poof is finished by the definition $\beta_t = \sqrt{2\lambda_t}$. ■

A.2 Proof of Lemma 8

Proof. Since δ_t is a one-hot vector, for all $j \in [N]$, there is a single index $j \in [0]$ for which $\delta_j = 1$ and $\delta_{j'} = 0$ for all $j' \neq j$. Besides, we have $\Pr(\delta_{i,j} = 1|\mathcal{F}_i) = p_{i,j}$. Hence, conditional on $\mathcal{F}_i$, the variance of $\xi^\top z_i$ can be written as follows. For simplicity, we denote $\mathbb{E}[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_i]$ below.

$$\begin{aligned}
&\sigma^2(\xi^\top z_i|\mathcal{F}_i) \\
&= \mathbb{E}\left[\left(\sum_{j=1}^N (\delta_{i,j} - p_{i,j})\xi^\top x_{i,j}\right)^2\right] - \mathbb{E}\left[\left(\sum_{j=1}^N (\delta_{i,j} - p_{i,j})\xi^\top x_{i,j}\right)\right]^2 \\
&= \mathbb{E}\left[\left(\sum_{j=1}^N (\delta_{i,j} - p_{i,j})\xi^\top x_{i,j}\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{j=1}^N\sum_{k=1}^N (\delta_{i,j}\xi^\top x_{i,j})(\delta_{i,k}\xi^\top x_{i,k})\right)^2\right] - 2\mathbb{E}\left[\sum_{j=1}^N \delta_{i,j}\xi^\top x_{i,j}\right]\left(\sum_{j=1}^N p_{i,j}\xi^\top x_{i,j}\right) + \left(\sum_{j=1}^N p_{i,j}\xi^\top x_{i,j}\right)^2 \\
&= \mathbb{E}\left[\sum_{j=1}^N \delta_{i,j}(\xi^\top x_{i,j})^2\right] - 2\left(\sum_{j=1}^N p_{i,j}\xi^\top x_{i,j}\right)^2 + \left(\sum_{j=1}^N p_{i,j}\xi^\top x_{i,j}\right)^2 \\
&= \sum_{j=1}^N p_{i,j}(\xi^\top x_{i,j})^2 - \left(\sum_{j=1}^N p_{i,j}\xi^\top x_{i,j}\right)^2 \leq \sum_{j=1}^N (\xi^\top x_{i,j})^2 = \|\xi\|_{Y_t}^2 - \|\xi\|_{Y_{t-1}}^2.
\end{aligned}$$

Then, note that S_{t-1} is $\mathcal{F}_t$-measurable, thus for all $t \geq 1$, we have

$$\mathbb{E}\left[\exp(\xi^\top S_t)|\mathcal{F}_t\right] = \exp(\xi^\top S_{t-1})\mathbb{E}\left[\exp(\xi^\top z_t)|\mathcal{F}_t\right].$$

Next we apply Lemma 12 to bound the term $\mathbb{E}\left[\exp(\xi^\top z_t)|\mathcal{F}_t\right]$ and we need to ensure $|\xi^\top z_t| \leq 1$. Since δ_t is an one-hot vector, let j be the index such that $\delta_{t,j} = 1$. Then, we have $\delta_{t,k} = 0$ for all $k \in [N]\backslash\{j\}$. Note $\|x_{t,j}\| \leq 1$ for all t, j and $\|\xi\| \leq 1/2$, thus we have

$$|\xi^\top z_t| \leq (1 - p_{t,j})|\xi^\top x_{t,j}| + \sum_{k\in[N]\backslash\{j\}} p_{t,k}|\xi^\top x_{t,k}| \leq \frac{1}{2}\left(1 + \sum_{j\in[N]} p_{t,j}\right) \leq 1.$$

Since $|\xi^\top z_t| \leq 1$, we can apply Lemma 12 and obtain

$$\mathbb{E}\left[\exp(\xi^\top S_t)|\mathcal{F}_t\right] = \exp(\xi^\top S_{t-1})\mathbb{E}\left[\exp(\xi^\top z_t)|\mathcal{F}_t\right] \leq \exp(\xi^\top S_{t-1})(1 + \sigma^2(\xi^\top z_t|\mathcal{F}_t)) \leq \exp(\xi^\top S_{t-1} + \sigma^2(\xi^\top z_t|\mathcal{F}_t)), \tag{16}$$

where the first inequality holds by Lemma 12 and the last inequality holds by $1 + x \leq e^x$.

Finally, we obtain

$$\begin{aligned}
\mathbb{E}\left[M_t(\xi)|\mathcal{F}_t\right] = \mathbb{E}\left[\exp(\xi^\top S_t - \|\xi\|_{Y_t})|\mathcal{F}_t\right] &= \mathbb{E}\left[\exp(\xi^\top S_t)|\mathcal{F}_t\right]\exp(-\|\xi\|_{Y_t}) \\
&\leq \exp(\xi^\top S_{t-1} + \sigma^2(\xi^\top z_t|\mathcal{F}_t)^2 - \|\xi\|^2_{Y_t}) \\
&\leq \exp(\xi^\top S_{t-1} + \|\xi\|^2_{Y_{t-1}}) \\
&= M_{t-1}(\xi),
\end{aligned}$$

where the first equality holds since Y_t is $\mathcal{F}_t$-measurable, the first inequality holds by (16) and the second inequality holds by $\sigma^2(\xi^\top z_t|\mathcal{F}_t)^2 \leq \|\xi\|^2_{Y_t} - \|\xi\|^2_{Y_{t-1}}$. This shows that $\{M_t(\xi)\}_{t=0}^\infty$ is a super martingale. ■

B Proof of Lemma 2

Proof. Recall the closed-form of $\theta_{k,h}$ is given by

$$\theta_{k,h} = \Lambda_{k,h}^{-1}\sum_{i=1}^{k-1}\sum_{s'\in\mathcal{S}_{h+1}} \delta_{s_{i,h+1}}(s')\phi(s' \mid s_{i,h}, a_{i,h}).$$

where $\Lambda_{k,h} = \sum_{i=1}^{k-1}\sum_{s'\in\mathcal{S}_{h+1}} \phi(s'|s_{i,h}, a_{i,h})\phi(s'|s_{i,h}, a_{i,h})^\top + \lambda_k I_d$. We decompose the closed form as follows.

$$\begin{aligned}
\theta_{k,h} &= \Lambda_{k,h}^{-1}\sum_{i=1}^{k-1}\sum_{s'\in\mathcal{S}_{h+1}} \delta_{s_{i,h+1}}(s')\phi(s' \mid s_{i,h}, a_{i,h}) \\
&= \Lambda_{k,h}^{-1}\sum_{i=1}^{k-1}\sum_{s'\in\mathcal{S}_{h+1}} (P_h(s' \mid s_{i,h}, a_{i,h}) + \varepsilon_{i,h}(s'))\phi(s' \mid s_{i,h}, a_{i,h}) \\
&= \Lambda_{k,h}^{-1}\sum_{i=1}^{k-1}\sum_{s'\in\mathcal{S}_{h+1}} P_h(s' \mid s_{i,h}, a_{i,h})\phi(s' \mid s_{i,h}, a_{i,h}) + \Lambda_{k,h}^{-1}\sum_{i=1}^{k-1}\sum_{s'\in\mathcal{S}_{h+1}} \varepsilon_{i,h}(s')\phi(s' \mid s_{i,h}, a_{i,h}) \\
&= \Lambda_{k,h}^{-1}\sum_{i=1}^{k-1}\sum_{s'\in\mathcal{S}_{h+1}} \phi_h(s' \mid s_{i,h}, a_{i,h})^\top\theta_h^*\phi(s' \mid s_{i,h}, a_{i,h}) + \Lambda_{k,h}^{-1}\sum_{i=1}^{k-1}\sum_{s'\in\mathcal{S}_{h+1}} \varepsilon_{i,h}(s')\phi(s' \mid s_{i,h}, a_{i,h}) \\
&= \Lambda_{k,h}^{-1}(\Lambda_{k,h} - \lambda_k I_d)\theta_h^* + \Lambda_{k,h}^{-1}\sum_{i=1}^{k-1}\sum_{s'\in\mathcal{S}_{h+1}} \varepsilon_{i,h}(s')\phi(s' \mid s_{i,h}, a_{i,h}) \\
&= \theta_h^* + \lambda_k\Lambda_{k,h}^{-1}\theta_h^* + \Lambda_{k,h}^{-1}\sum_{i=1}^{k-1}\sum_{s'\in\mathcal{S}_{h+1}} \varepsilon_{i,h}(s')\phi(s' \mid s_{i,h}, a_{i,h}).
\end{aligned}$$

Rearranging terms, we obtain for all $\zeta \in (0,1)$, with probability at least $1-\zeta/H$, it holds that

$$\begin{aligned}\|\theta_{k,h}-\theta_h^*\|_{\Lambda_{k,h}} &= \lambda_k\|\theta_h^*\|_{\Lambda_{k,h}} + \Big\|\sum_{i=1}^{k-1}\sum_{s'\in\mathcal{S}_{h+1}} \varepsilon_{i,h}(s')\phi(s' \mid s_{i,h},a_{i,h})\Big\|_{\Lambda_{k,h}} \\ &\leq \sqrt{\lambda_k}B + \frac{\sqrt{\lambda_k}}{4} + \frac{4}{\sqrt{\lambda_k}}\log\left(\frac{2^d\det(\Lambda_{k,h})^{\frac{1}{2}}\lambda_k^{-\frac{d}{2}}}{\zeta/H}\right) \\ &\leq \sqrt{\lambda_k}(B+\frac{1}{4}) + \frac{4}{\sqrt{\lambda_k}}\left(\log\left(\frac{H}{\zeta}\right)+\frac{d}{2}\log\left(4+\frac{4Sk}{\lambda_k d}\right)\right),\end{aligned}$$

where the first inequality holds by the self-normalized concentration of Lemma 1, and the last inequality holds by the determinant-trace inequality in Lemma 16. This shows with probability at least $1-\zeta/H$, it holds that $\theta_h^* \in \mathcal{C}_{k,h}$. Applying a union bound over $h=1,\ldots,H$ finishes the proof. ■

C Proof of Lemma 3

In this section, we present the proof of Lemma 3.

C.1 Main Proof

Proof. By Lemma 4, for any $(s,a)\in\mathcal{S}_m\times\mathcal{A}, m\in[H]$, we have

$$\sum_{s'\in\mathcal{S}_{m+1}} \epsilon_{k,m}(s' \mid s,a) = \sum_{s'\in\mathcal{S}_{m+1}} |\widehat{P}_{k,m}(s'|s,a) - P_m(s'|s,a)| \leq 2 \wedge \sum_{s'\in\mathcal{S}_{m+1}} \beta_k\|\phi(s' \mid s,a)\|_{\Lambda_{k,m}^{-1}},$$

where the last inequality holds by the fact $\|\widehat{P}_{k,m}(\cdot \mid s,a)-P_m(\cdot \mid s,a)\|_1 \leq \|\widehat{P}_{k,m}(\cdot \mid s,a)\|_1 + \|P_m(\cdot \mid s,a)\|_1 = 2$.

Then, with probability at least $1-\zeta$, it holds that

$$\begin{aligned}&\sum_{k=1}^K\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} |q_k^s(s,a)-q_k(s,a)| \\ &\leq 2S\sum_{k=1}^K\sum_{m=1}^H\sum_{s'\in\mathcal{S}_{m+1}} \epsilon_{k,h}(s' \mid s_{k,m},a_{k,m}) + 4S^2\log\left(\frac{H}{\zeta}\right) \\ &\leq 4S\sum_{k=1}^K\sum_{m=1}^H \beta_k(1\wedge\sum_{s'\in\mathcal{S}_{m+1}}\|\phi(s' \mid s_{k,m},a_{k,m})\|_{\Lambda_{k,m}^{-1}}) + 4S^2\log\left(\frac{H}{\zeta}\right) \\ &\leq 4\beta_K S\sum_{m=1}^H\sum_{k=1}^K(1\wedge\sum_{s'\in\mathcal{S}_{m+1}}\|\phi(s' \mid s_{k,m},a_{k,m})\|_{\Lambda_{k,m}^{-1}}) + 4S^2\log\left(\frac{H}{\zeta}\right) \\ &\leq 2\beta_K S\sum_{m=1}^H\sqrt{KS_{m+1}\sum_{k=1}^K\left(1\wedge\sum_{s'\in\mathcal{S}_{m+1}}\|\phi(s' \mid s_{k,m},a_{k,m})\|^2_{\Lambda_{k,m}^{-1}}\right)} + 4S^2\log\left(\frac{H}{\zeta}\right) \\ &\leq 2\beta_K S\sum_{m=1}^H\sqrt{KS_{m+1}d\log\left(\lambda_{K+1}+\frac{KS_{m+1}}{d}\right)} + 4S^2\log\left(\frac{H}{\zeta}\right) \\ &\leq 2\beta_K S\sqrt{KH\sum_{m=1}^H S_{m+1}d\log\left(\lambda_{K+1}+\frac{KS}{d}\right)} + 4S^2\log\left(\frac{H}{\zeta}\right) \\ &\leq \mathcal{O}\left((d\sqrt{HS^3K}+S^2)\log^2\left(\frac{dKS}{\zeta}\right)\right) \\ &\leq \mathcal{O}\left(d\sqrt{HS^3K}\log^2\left(\frac{dKS}{\zeta}\right)\right),\end{aligned}$$

where the first inequality follows from Lemma 5, the fourth and sixth inequality holds by the Cauchy-Schwarz inequality, the fifth inequality holds by the specifically designed elliptical potential lemma in Lemma 17, the second last inequality holds by $\lambda_{K+1} = d\log((K+1)S)$ and $\beta_K = \mathcal{O}(\sqrt{d\log(KS)})$ and the last bound holds by $S \le K$ (otherwise the bound $\sqrt{HSAK}$ becomes vacuous). This completes the proof. ■

C.2 Proof of Lemma 4

Proof. This lemma was first proved in Zhao et al. (2023a, Lemma 3). We present their proof for self-containedness.

By the definition of linear mixture MDPs, for all $k \in [K]$, $h \in [H]$ and $\forall (s,a,s') \in \mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}$, we have

$$\begin{aligned}
\left|\widehat{P}_{k,h}(s' \mid s,a) - P_h(s' \mid s,a)\right| &= \left|\phi(s' \mid s,a)^\top (\widehat{\theta}_{k,h} - \theta_h^*)\right| \\
&\le \|\phi(s' \mid s,a)\|_{\Lambda_{k,h}^{-1}} \|\widehat{\theta}_{k,h} - \theta_h^*\|_{\Lambda_{k,h}} \\
&\le \beta_k \|\phi(s' \mid s,a)\|_{\Lambda_{k,h}^{-1}} \\
&\le 1 \wedge \beta_k \|\phi(s' \mid s,a)\|_{\Lambda_{k,h}^{-1}},
\end{aligned}$$

where the first inequality follows from the Holder's inequality, the second inequality holds by Lemma 2, and the last inequality follows from the fact that $|\widehat{P}_{k,h}(s' \mid s,a) - P_h(s' \mid s,a)| \le 1$. This completes the proof. ■

C.3 Proof of Lemma 5

Proof. The main proof is similar to the proof of Zhao et al. (2023a, Lemma 2). Let $q_k^s = q^{P_k^s, \pi_k}$ for simplicity, and define $q(s) = \sum_{a \in \mathcal{A}} q(s,a)$. For any q and any (s,a), we have

$$\begin{aligned}
&q(s,a) = q(s)\pi^q(a \mid s) \\
&= \pi^q(s,a) \sum_{s' \in \mathcal{S}_{h(s)-1}} q(s') \sum_{a' \in \mathcal{A}} \pi^q(a' \mid s') P^q(s \mid s', a') \\
&= \pi^q(s \mid a) \sum_{\{s_i,a_i\}_{i=1}^{h(s)-1} \in \prod_{i=1}^{h(s)-1} \mathcal{S}_i \times \mathcal{A}} \prod_{h=1}^{h(s)-1} \pi^q(a_h \mid s_h) \prod_{h=1}^{h(s)-1} P^q(s_{h+1} \mid s_h, a_h),
\end{aligned}$$

where the last equality holds by expressing $q(s_{i+1})$ using $q(s_i)$ recursively for $i = h(s)-1, \ldots, 1$. In the following, we drop the superscript $\prod_{i=1}^{h(s)-1} \mathcal{S}_i \times \mathcal{A}$ of $\{s_i,a_i\}_{i=1}^{h(s)-1}$ for simplicity. Then, we have

$$\begin{aligned}
&|q_k^s(s,a) - q_k(s,a)| \\
&= \pi_k(s \mid a) \sum_{\{s_i,a_i\}_{i=1}^{h(s)-1}} \prod_{h=1}^{h(s)-1} \pi_k(a_h \mid s_h) \left(\prod_{h=1}^{h(s)-1} P_k^s(s_{h+1} \mid s_h, a_h) - \prod_{h=1}^{h(s)-1} P(s_{h+1} \mid s_h, a_h) \right)
\end{aligned}$$

Further, we decompose the difference of the transition as follows.

$$\begin{aligned}
&\prod_{h=1}^{h(s)-1} P_k^s(s_{h+1} \mid s_h, a_h) - \prod_{h=1}^{h(s)-1} P(s_{h+1} \mid s_h, a_h) \\
&= \prod_{h=1}^{h(s)-1} P_k^s(s_{h+1} \mid s_h, a_h) - \prod_{h=1}^{h(s)-1} P(s_{h+1} \mid s_h, a_h) \pm \sum_{m=1}^{h(s)-1} \prod_{h=1}^{m-1} P(s_{h+1} \mid s_h, a_h) \prod_{h=m}^{h(s)-1} P_k^s(s_{h+1} \mid s_h, a_h) \\
&= \sum_{m=1}^{h(s)-1} \left(P_k^s(s_{m+1} \mid s_m, a_m) - P(s_{m+1} \mid s_m, a_m)\right) \prod_{h=1}^{m-1} P(s_{h+1} \mid s_h, a_h) \prod_{h=m+1}^{h(s)-1} P_k^s(s_{h+1} \mid s_h, a_h) \\
&\le \sum_{m=1}^{h(s)-1} \epsilon_{k,m}(s_{m+1} \mid s_m, a_m) \prod_{h=1}^{m-1} P(s_{h+1} \mid s_h, a_h) \prod_{h=m+1}^{h(s)-1} P_k^s(s_{h+1} \mid s_h, a_h),
\end{aligned}$$

where $\epsilon_{k,h}(s' \mid s,a) = 1 \wedge \beta_k \|\phi(s' \mid s,a)\|_{\Lambda_{k,h}^{-1}}$. Therefore, we have

$$
\begin{aligned}
&|q_k^s(s,a) - q_s(s,a)| \\
&\leq \pi_k(s \mid a) \sum_{\{s_i,a_i\}_{i=1}^{h(s)-1}} \prod_{h=1}^{h(s)-1} \pi_k\left(a_h \mid x_h\right) \sum_{m=1}^{h(s)-1} \epsilon_{k,m}\left(x_{m+1} \mid x_m, a_m\right) \prod_{h=1}^{m-1} P\left(s_{h+1} \mid s_h, a_h\right) \prod_{h=m+1}^{h(s)-1} P_k^s\left(s_{h+1} \mid s_h, a_h\right) \\
&= \sum_{m=1}^{h(s)-1} \sum_{\{s_i,a_i\}_{i=1}^{h(s)-1}} \epsilon_{k,m}\left(s_{m+1} \mid s_m, a_m\right) \left(\pi_k\left(a_m \mid s_m\right) \prod_{h=1}^{m-1} \pi_k\left(a_h \mid s_h\right) P\left(s_{h+1} \mid s_h, a_h\right)\right) \\
&\qquad\qquad \cdot \left(\pi_k(s \mid a) \prod_{h=m+1}^{h(s)-1} \pi_k\left(a_h \mid x_h\right) P_k^s\left(s_{h+1} \mid s_h, a_h\right)\right) \\
&= \sum_{m=1}^{h(s)-1} \sum_{s_m,a_m,s_{m+1}} \epsilon_{k,m}\left(s_{m+1} \mid s_m, a_m\right) \left(\sum_{\{s_i,a_i\}_{i=1}^{m-1}} \pi_k\left(a_m \mid x_m\right) \prod_{h=1}^{m-1} \pi_k\left(a_h \mid s_h\right) P\left(s_{h+1} \mid s_h, a_h\right)\right) \\
&\qquad\qquad \cdot \left(\sum_{a_{m+1}} \sum_{\{s_i,a_i\}_{i=m+2}^{h(s)-1}} \pi_k(s \mid a) \prod_{h=m+1}^{h(s)-1} \pi_k\left(a_h \mid s_h\right) P_k^s\left(s_{h+1} \mid s_h, a_h\right)\right) \\
&= \sum_{m=1}^{h(s)-1} \sum_{s_m,a_m,s_{m+1}} \epsilon_{k,m}\left(s_{m+1} \mid s_m, a_m\right) q_k\left(s_m, a_m\right) \pi_k(a \mid s) q_k^s\left(s \mid s_{m+1}\right) \\
&\leq \pi_k(a \mid s) \sum_{m=1}^{h(s)-1} \sum_{s_m,a_m,s_{m+1}} \epsilon_{k,m}\left(s_{m+1} \mid s_m, a_m\right) q_k\left(s_m, a_m\right),
\end{aligned}
$$

where the last inequality holds by $q_k^s\left(s \mid s_{m+1}\right) \leq 1$.

Let $w_m = (s_m, a_m, s_{m+1})$ to simplify the notation. Then, summing over $k \in [K]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$
\begin{aligned}
&\sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} |q_k^s(s,a) - q_k(s,a)| \\
&\leq \sum_{k,s,a} \pi_k(a \mid s) \sum_{m=1}^{h(s)-1} \sum_{w_m} \epsilon_{k,m}\left(s_{m+1} \mid s_m, a_m\right) q_k\left(s_m, a_m\right) \\
&= \sum_{k} \sum_{h\leq H} \sum_{m=1}^{h-1} \sum_{w_m} \epsilon_{k,m}\left(s_{m+1} \mid s_m, a_m\right) q_k\left(s_m, a_m\right) \sum_{(s,a)\in\mathcal{S}_h\times\mathcal{A}} \pi_k(a \mid s) \\
&= \sum_{1\leq m<h\leq H} \sum_{k,w_m} \epsilon_{k,m}\left(s_{m+1} \mid s_m, a_m\right) q_k\left(s_m, a_m\right) |\mathcal{S}_h| \\
&\leq S \sum_{1\leq m\leq H} \sum_{k,w_m} \epsilon_{k,m}\left(s_{m+1} \mid s_m, a_m\right) q_k\left(s_m, a_m\right)
\end{aligned} \tag{17}
$$

Then, we focus on $\sum_{k,w_m} \epsilon_{k,m}\left(s_{m+1} \mid s_m, a_m\right) q_k\left(s_m, a_m\right)$ with a fixed m at first:

$$
\begin{aligned}
&\sum_{k,w_m} \epsilon_{k,m}\left(s_{m+1} \mid s_m, a_m\right) q_k\left(s_m, a_m\right) \\
&= \underbrace{\sum_{k,w_m} \mathbb{I}_k\{s_m, a_m\} \epsilon_{k,m}\left(s_{m+1} \mid s_m, a_m\right)}_{\textsc{Term-I}} + \underbrace{\sum_{k,w_m} S_{m+1}\left(\frac{q_k(s_m,a_m)}{S_{m+1}} - \frac{\mathbb{I}_k(s_m,a_m)}{S_{m+1}}\right) \epsilon_{k,m}\left(s_{m+1} \mid s_m, a_m\right)}_{\textsc{Term-II}}
\end{aligned} \tag{18}
$$

For TERM-I, since $\mathbb{I}_k(s,a)$ is the indicator whether the pair (s,a) is visited in episode k, thus we have

$$\sum_{k,w_m} \mathbb{I}_k\{s_m,a_m\}\epsilon_{k,m}\left(s_{m+1}\mid s_m,a_m\right) = \sum_{k=1}^{K}\sum_{s'\in\mathcal{S}_{m+1}} \epsilon_{k,h}(s'\mid s_{k,m},a_{k,m}) \tag{19}$$

To bound TERM-II, we first use Lemma 13 to build the connection between TERM-I and TERM-II. Let

$$Y_{k,m} = \sum_{w_m}\left(\frac{q_k\left(s_m,a_m\right)}{S_{m+1}} - \frac{\mathbb{I}_k\left\{s_m,a_m\right\}}{S_{m+1}}\right)\epsilon_{k,m}\left(s_{m+1}\mid s_m,a_m\right).$$

It is easy to verify that $Y_{k,m}\le 1$. Let $o_{i,j}=(s_{i,j},a_{i,j},\ell_i(s_{i,j},a_{i,j}))$ be the observations in episode i, we denote $\mathcal{F}_{k,h}$ the σ-algebra generated by $\{o_{i,j}\}_{i=1,j=1}^{k,h}$. Then, we have

$$\begin{aligned}
\mathbb{E}_{k-1,H}\left[Y_{k,m}^2\right] \le & \frac{\mathbb{E}_{k-1,H}\left[\left(\sum_{w_m}\mathbb{I}_k\left(s_m,a_m\right)\epsilon_{k,m}\left(s_{m+1}\mid s_m,a_m\right)\right)^2\right]}{S_{m+1}^2}\\
= & \frac{\mathbb{E}_{k-1,H}\left[\sum_{w_m}\mathbb{I}_k\left(s_m,a_m\right)\epsilon_{k,m}\left(s_{m+1}\mid s_m,a_m\right)^2\right]}{S_{m+1}^2}\\
\le & \frac{\sum_{w_m} q_k\left(s_m,a_m\right)\epsilon_{k,m}\left(s_{m+1}\mid s_m,a_m\right)}{S_{m+1}}
\end{aligned}$$

where the equality follows from the fact that $\mathbb{I}_k(s_m,a_m)\,\mathbb{I}_k(s'_m,a'_m)$ for $s_m\neq s'_m$, and the last inequality holds by $\epsilon_{k,m}\left(s_{m+1}\mid s_m,a_m\right)\le 1$ and $\epsilon_{k,m}$ is $\mathcal{F}_{k-1,H}$-measurable. Then, by choosing $\lambda=1/2$ in Lemma 13, with probability at least $1-\zeta/H$, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{w_m}\left(\frac{q_k\left(s_m,a_m\right)}{S_{m+1}} - \frac{\mathbb{I}_k\left\{s_m,a_m\right\}}{S_{m+1}}\right)\epsilon_{k,m}\left(s_{m+1}\mid s_m,a_m\right)\\
\le\ & \frac{1}{2S_{m+1}}\sum_{k=1}^{K}\sum_{w_m} q_k\left(s_m,a_m\right)\epsilon_{k,m}\left(s_{m+1}\mid s_m,a_m\right) + 2\log(H/\zeta)
\end{aligned}$$

By applying with a union bound over $m=1,\ldots,H$, we have with probability at least $1-\zeta$, it holds that

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{w_m}\left(q_k\left(s_m,a_m\right) - \mathbb{I}_k\left\{s_m,a_m\right\}\right)\epsilon_{k,m}\left(s_{m+1}\mid s_m,a_m\right)\\
\le\ & \frac{1}{2}\sum_{k=1}^{K}\sum_{w_m} q_k\left(s_m,a_m\right)\epsilon_{k,m}\left(s_{m+1}\mid s_m,a_m\right) + 2S_{m+1}\log(H/\zeta).
\end{aligned}$$

This shows that

$$\text{TERM-II} \le \sum_{k=1}^{K}\sum_{w_m}\mathbb{I}_k\left\{s_m,a_m\right\}\epsilon_{k,m}\left(s_{m+1}\mid s_m,a_m\right) + 4S_{m+1}\log(H/\zeta) \le \text{TERM-I} + 4S_{m+1}\log(H/\zeta). \tag{20}$$

Combining (17) and (18), we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|q_k^s(s,a) - q_k(s,a)\right|\\
\le\ & S\sum_{1\le m\le H}\left(\text{TERM-I} + \text{TERM-II}\right)\\
\le\ & S\sum_{1\le m\le H}\left(2\text{TERM-I} + 4S_{m+1}\log(H/\zeta)\right)\\
\le\ & 2S\sum_{k=1}^{K}\sum_{m=1}^{H}\sum_{s'\in\mathcal{S}_{m+1}}\epsilon_{k,h}(s'\mid s_{k,m},a_{k,m}) + 4S^2\log(H/\zeta),
\end{aligned}$$

where the second inequality holds by (20) and the last inequality holds by (19). This completes the proof. ∎

D Proof of Theorem 1

In this section, we present the proof of Theorem 1.

D.1 Main Proof

Proof. Define the occupancy measure under the true transition P and policy π_k as $q_k = q^{P,\pi_k}$. Then, the regret can be written as $\text{Reg} = \sum_{k=1}^K \langle q_k - q^*, \ell_k \rangle$. As in Section 5.3, we decompose the regret as the follows:

$$\text{Reg}(K) \leq \underbrace{\sum_{k=1}^K \langle \widehat{q}_k - q^*, \widehat{\ell}_k \rangle}_{\text{REGRET}} + \underbrace{\sum_{k=1}^K \langle q_k - \widehat{q}_k, \ell_k \rangle}_{\text{ERROR}} + \underbrace{\sum_{k=1}^K \langle \widehat{q}_k, \ell_k - \widehat{\ell}_k \rangle}_{\text{BIAS-I}} + \underbrace{\sum_{k=1}^K \langle q^*, \widehat{\ell}_k - \ell_k \rangle}_{\text{BIAS-II}}.$$

The bounds of ERROR and BIAS-I term are shown in Lemma 6 and Lemma 7 of Section 5, respectively. We bound the REGRET and BIAS-II terms below.

Bounding BIAS-II **Term.** For this term, we present the following lemma, whose proof is in Appendix D.2.

Lemma 10. *For any $\zeta \in (0, 1)$, with probability at least $1 - 2\zeta$, VLSUOB-REPS algorithm ensures that*

$$\text{BIAS-II} \leq \mathcal{O}\left(\frac{H \log(SA/\zeta)}{\gamma}\right).$$

Bounding REGRET **Term.** For this term, we present the following lemma, whose proof is in Appendix D.3.

Lemma 11. *For any $\zeta \in (0, 1)$, with probability at least $1 - 2\zeta$, VLSUOB-REPS algorithm ensures that*

$$\text{REGRET} \leq \mathcal{O}\left(\frac{H \log(SA/\zeta)}{\eta} + \eta SAK + \frac{\eta H \log(H/\zeta)}{\gamma}\right).$$

Combining Lemma 6, Lemma 7, Lemma 10 and Lemma 11 finishes the proof. ∎

D.2 Proof of Lemma 10

Proof. For some $(s, a) \in \mathcal{S} \times \mathcal{A}$, using $\alpha_k(s', a') = 2\gamma \mathbb{I}\{(s', a') = (s, a)\}$, we have with probability at least $1 - \frac{\zeta}{SA}$,

$$\sum_{k=1}^K \left(\widehat{\ell}_k(s, a) - \frac{q_k(s, a)}{u_k(s, a)} \ell_k(s, a)\right) \leq \frac{1}{2\gamma} \log\left(\frac{SA}{\zeta}\right)$$

By using a union bound, the above inequality holds for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ simultaneously with probability at least $1 - \zeta$. Further, under the event that $\theta_h^* \in \mathcal{C}_{k,h}$, we have $q_k(s, a) \leq u_k(s, a)$, which implies that

$$\begin{aligned}
\sum_{k=1}^K \left\langle q^*, \widehat{\ell}_k - \ell_k \right\rangle &\leq \sum_{k,s,a} q^*(s, a) \ell_k(s, a) \left(\frac{q_k(s, a)}{u_k(s, a)} - 1\right) + \sum_{s,a} \frac{q^*(s, a) \log \frac{SA}{\zeta}}{2\gamma} \\
&= \sum_{k,s,a} q^*(s, a) \ell_k(s, a) \left(\frac{q_k(s, a)}{u_k(s, a)} - 1\right) + \frac{H \log \frac{SA}{\zeta}}{2\gamma} \\
&\leq \frac{H \log \frac{SA}{\zeta}}{2\gamma}.
\end{aligned}$$

The proof is concluded by applying the union bound again. ∎

D.3 Proof of Lemma 11

Proof. The update procedure in (12) can be written as the following two-step procedure.

$$\widetilde{q}_{k+1} = \operatorname*{arg\,min}_{q \in \mathbb{R}_+^{SA}} \eta \left\langle q, \widehat{\ell}_k \right\rangle + D_\psi\left(q, \widehat{q}_k\right),$$
$$\widehat{q}_{k+1} = \operatorname*{arg\,min}_{q \in \Delta(\mathcal{P}_{k+1})} D_\psi\left(q, \widetilde{q}_{k+1}\right),$$

The closed form of $\widetilde{q}_{k+1}$ is given by $\widetilde{q}_{k+1}(s,a) = \widehat{q}_{k+1}(s,a)\exp(-\eta\widehat{\ell}_k(s,a))$. Then, we have

$$\begin{aligned}\left\langle \widehat{q}_k - q^*, \widehat{\ell}_k \right\rangle &= \frac{1}{\eta}\left(D_\psi\left(q^*, \widehat{q}_k\right) + D_\psi\left(\widehat{q}_k, \widetilde{q}_{k+1}\right) - D_\psi\left(q^*, \widetilde{q}_{k+1}\right)\right)\\ &\le \frac{1}{\eta}\left(D_\psi\left(q^*, \widehat{q}_k\right) + D_\psi\left(\widehat{q}_k, \widetilde{q}_{k+1}\right) - D_\psi\left(q^*, \widehat{q}_{k+1}\right)\right),\end{aligned}$$

where the equality holds by the three-point equality, and the inequality holds by the generalized Pythagorean theorem. Then, summing over $k \in [K]$ and using the telescoping argument, we have

$$\sum_{k=1}^K \left\langle \widehat{q}_k - q^*, \widehat{\ell}_k \right\rangle \le \frac{1}{\eta}\left(D_\psi\left(q^*, \widehat{q}_1\right) - D_\psi\left(q^*, \widehat{q}_{K+1}\right) + \sum_{k=1}^K D_\psi\left(\widehat{q}_k, \widetilde{q}_{k+1}\right)\right).$$

The first two terms can be rewritten as

$$D_\psi\left(q^*, \widehat{q}_1\right) - D_\psi\left(q^*, \widehat{q}_{K+1}\right) = \sum_{h=1}^H \sum_{s \in \mathcal{S}_h} \sum_{a \in \mathcal{A}} q^*(s,a) \log \frac{\widehat{q}_{K+1}(s,a)}{\widehat{q}_1(s,a)} \le \sum_{h=1}^H \sum_{s \in \mathcal{S}_h} \sum_{a \in \mathcal{A}} q^*(s,a) \log\left(S_h A\right) \le H\log(SA).$$

It remains to bound the last term.

$$\begin{aligned}D_\psi\left(\widehat{q}_k, \widetilde{q}_{k+1}\right) &= \sum_{h=1}^H \sum_{s \in \mathcal{S}_h} \sum_{a \in \mathcal{A}} \left(\eta\widehat{q}_k\left(s,a\right)\widehat{\ell}_k(s,a) - \widehat{q}_k\left(s,a\right) + \widehat{q}_k\left(s,a\right)\exp\left(-\eta\widehat{\ell}_k(s,a)\right)\right)\\ &\le \eta^2 \sum_{h=1}^H \sum_{s \in \mathcal{S}_h} \sum_{a \in \mathcal{A}} \widehat{q}_k\left(s,a\right)\widehat{\ell}_k(s,a)^2 = \eta^2 \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \widehat{q}_k(s,a)\widehat{\ell}_k(s,a)^2,\end{aligned}$$

where the inequality is due to the fact that $e^{-x} \le 1 - x + x^2$ for all $x \ge 0$.

Note that due to the definition of $\widehat{\ell}_k(s,a)$, we have

$$\widehat{q}_k(s,a)\widehat{\ell}_k(s,a)^2 = \frac{\widehat{q}_k(s,a)\ell_k(s,a)\mathbb{I}_k\{s,a\}}{u_k(s,a)+\gamma}\widehat{\ell}_k(s,a) \le \widehat{\ell}_k(s,a),$$

which is due to the fact that $\widehat{q}_k(s,a) \le u_k(s,a)$ and $\ell_k(s,a)\mathbb{I}_k\{s,a\} \le 1$. Furthermore, using Lemma 13 by setting $\alpha_k(s,a) = 2\gamma$, with probability at least $1-\zeta$, we have

$$\sum_{k,s,a} \widehat{q}_k(s,a)\widehat{\ell}_k(s,a)^2 \le \sum_{k,s,a} \frac{q_k(s,a)}{u_k(s,a)}\ell_k(s,a) + \frac{H\log(\frac{H}{\zeta})}{2\gamma} \le SAK + \frac{H\log(\frac{H}{\zeta})}{2\gamma}$$

where the last inequality comes from that $q_k(s,a) \le u_k(s,a)$ and $\ell_k(s,a) \le 1$.

Applying a union bound over the above bounds, with probability at least $1-2\zeta$, we have

$$\sum_{k=1}^K \left\langle \widehat{q}_k - q^*, \widehat{\ell}_k \right\rangle \le \frac{H\log(SA)}{\eta} + \eta SAK + \frac{\eta H\log(H/\zeta)}{\gamma}.$$

This finishes the proof. ∎

E Supporting Lemmas

In this section, we present some supporting lemmas, which are useful in our proofs.

First, we introduce the following two lemmas, which are used in the analysis of super martingale.

Lemma 12 (Lemma 7 of Faury et al. (2020)). *Let ϵ be a centered random variable of variance σ^2 such that $|\epsilon| \leq 1$ almost surely. Then for all $\lambda \in [-1, 1]$ we have $\mathbb{E}[\exp(\lambda\epsilon)] \leq 1 + \lambda^2\sigma^2$.*

Lemma 13 (Theorem 1 of Beygelzimer et al. (2011)). *Let $Y_1, \ldots, Y_K$ be a martingale difference sequence with respect to a filtration $\mathcal{F}_1, \ldots, \mathcal{F}_K$. Suppose that $|Y_k| \leq R$ for all $k \in [K]$. Then, for any $\zeta \in (0, 1)$ and $\lambda \in [0, 1/R]$, with probability at least $1 - \zeta$, we have $\sum_{k=1}^K Y_k \leq \lambda \sum_{k=1}^K \mathbb{E}[Y_k^2 | \mathcal{F}_{k-1}] + \frac{\log(1/\zeta)}{\lambda}$.*

Then, we introduce the lemma that guarantees the biased loss estimator is close to the true loss function.

Lemma 14 (Lemma 11 of Jin et al. (2020a)). *For any sequence of functions $\alpha_1, \ldots, \alpha_K$ such that $\alpha_k \in [0, 2\gamma]^{S \times A}$ if $\mathcal{F}_{k-1,H}$-measurable for all $k \in [k]$, with probability at least $1 - \zeta$, we have*

$$\sum_{k=1}^{K} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \alpha_k(s, a) \left(\widehat{\ell}_k(s, a) - \frac{q_k(s, a)}{u_k(s, a)} \ell_k(s, a) \right) \leq H \log \left(\frac{H}{\zeta} \right).$$

Next, we present the self-normalized concentration and determinant-trace lemma of Abbasi-Yadkori et al. (2011).

Lemma 15 (Theorem 1 of Abbasi-Yadkori et al. (2011)). *Let $\{\mathcal{F}_t\}_{t=0}^\infty$ be a filtration. Let $\{\eta_t\}_{t=1}^\infty$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally zero-mean R-sub-Gaussian for $R \geq 0$ i.e. $\forall \lambda \in \mathbb{R}, \mathbb{E}\left[e^{\lambda\eta_t} \mid \mathcal{F}_{t-1}\right] \leq \exp\left(\lambda^2 R^2/2\right)$. Let $\{X_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\mathcal{F}_{t-1}$-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \geq 0$, define $\bar{V}_t = V + \sum_{s=1}^t X_s X_s^\top$ and $S_t = \sum_{s=1}^t \eta_s X_s$. Then, for any $\zeta > 0$, with probability at least $1 - \zeta$, for all $t \geq 0$,*

$$\|S_t\|_{\bar{V}_t^{-1}}^2 \leq 2R^2 \log \left(\frac{\det\left(\bar{V}_t\right)^{1/2} \det(V)^{-1/2}}{\zeta} \right).$$

Lemma 16 (Lemma 10 of Abbasi-Yadkori et al. (2011)). *Suppose $x_1, \ldots, x_t \in \mathbb{R}^d$ and for any $1 \leq s \leq t$, $\|x_s\|_2 \leq L$. Let $V_t = \lambda I_d + \sum_{s=1}^t x_s x_s^\top$ for some $\lambda \geq 0$. Then, for any $1 \leq s \leq t$, we have $\det(V_t) \leq (\lambda + tL^2/d)^d$.*

Finally, we introduce the generalized elliptical potential lemma, which is designed specifically for our analysis.

Lemma 17 (Generalized elliptical potential lemma). *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_t \in \mathbb{R}^{N \times d}$ and for any $1 \leq s \leq t$, $\|x_{s,i}\|_2 \leq L$. Let $\Lambda_t = \lambda_t I_d + \sum_{s=1}^{t-1} \sum_{i=1}^N \mathbf{x}_{s,i} \mathbf{x}_{s,i}^\top$ with $\lambda_t \geq \lambda_{t-1}$ and $\lambda_1 = 1$. Then, for any $1 \leq s \leq t$, we have*

$$\sum_{s=1}^{t} \left(1 \wedge \sum_{i=1}^{N} \|\mathbf{x}_{s,i}\|_{\Lambda_s^{-1}} \right) \leq 2d \log \left(\lambda_{t+1} + \frac{tNL^2}{d} \right).$$

Proof. By the definition of Λ_t, we have

$$\det(\Lambda_{t+1}) = \det \left(\Lambda_t + \sum_{i=1}^{N} \mathbf{x}_{t,i} \mathbf{x}_{t,i}^\top + (\lambda_{t+1} - \lambda_t) I_d \right) \geq \det \left(\Lambda_t + \sum_{i=1}^{N} \mathbf{x}_{t,i} \mathbf{x}_{t,i}^\top \right) = \det(\Lambda_t) \left(1 + \sum_{i=1}^{N} \|\mathbf{x}_{t,i}\|_{\Lambda_t^{-1}}^2 \right),$$

where the inequality holds by the fact that $\lambda_{t+1} \geq \lambda_t$. Taking log from both sides and summing from $s = 1$ to t:

$$\sum_{s=1}^{t} \log \left(1 + \sum_{i=1}^{N} \|\mathbf{x}_{s,i}\|_{\Lambda_s^{-1}}^2 \right) = \log \left(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)} \right) \leq d \log(\lambda_{t+1} + \frac{tNL^2}{d})$$

where the last inequality holds by the determinant-trace inequality in Lemma 16. For any a such that $0 \leq a \leq 1$, it holds that $a \leq 2\log(1 + a)$. Thus, we have

$$\sum_{s=1}^{t} \left(1 \wedge \sum_{i=1}^{N} \|\mathbf{x}_{s,i}\|_{\Lambda_s^{-1}} \right) \leq 2 \sum_{s=1}^{t} \log \left(1 + \sum_{i=1}^{N} \|\mathbf{x}_{s,i}\|_{\Lambda_s^{-1}}^2 \right) \leq 2d \log \left(\lambda_{t+1} + \frac{tNL^2}{d} \right).$$

This completes the proof. ∎